Association analysis identifies 65 new breast cancer risk loci.

Kyriaki Michailidou[1-3], Sara Lindström[3-5], Joe Dennis[1,3], Jonathan Beesley[3,6], Shirley Hui[3,7], Siddhartha Kar[3,8], Audrey Lemaçon[9], Penny Soucy[9], Dylan Glubb[6], Asha Rostamianfar[7], Manjeet K. Bolla[1], Qin Wang[1], Jonathan Tyrer[8], Ed Dicks[8], Andrew Lee[1], Zhaoming Wang[10,11], Jamie Allen[1], Renske Keeman[12], Ursula Eilber[13], Juliet D. French[6], Xiao Qing Chen[6], Laura Fachal[8], Karen McCue[6], Amy E. McCart Reed[14], Maya Ghoussaini[8], Jason Carroll[15], Xia Jiang[5], Hilary Finucane[5,16], Marcia Adams[17], Muriel A. Adank[18], Habibul Ahsan[19], Kristiina Aittomäki[20], Hoda Anton-Culver[21], Natalia N. Antonenkova[22], Volker Arndt[23], Kristan J. Aronson[24], Banu Arun[25], Paul L. Auer[26,27], François Bacot[28], Myrto Barrdahl[13], Caroline Baynes[8], Matthias W. Beckmann[29], Sabine Behrens[13], Javier Benitez[30,31], Marina Bermisheva[32], Leslie Bernstein[33], Carl Blomqvist[34], Natalia V. Bogdanova[22,35,36], Stig E. Bojesen[37-39], Bernardo Bonanni[40], Anne-Lise Børresen-Dale[41], Judith S. Brand[42], Hiltrud Brauch[43-45], Paul Brennan[46], Hermann Brenner[23,45,47], Louise Brinton[48], Per Broberg[49], Ian W. Brock[50], Annegien Broeks[12], Angela Brooks-Wilson[51,52], Sara Y. Brucker[53], Thomas Brüning[54], Barbara Burwinkel[55,56], Katja Butterbach[23], Qiuyin Cai[57], Hui Cai[57], Trinidad Caldés[58], Federico Canzian[59], Angel Carracedo[60,61], Brian D. Carter[62], Jose E. Castelao[63], Tsun L. Chan[64,65], Ting-Yuan David Cheng[66], Kee Seng Chia[67], Ji-Yeob Choi[68,69], Hans Christiansen[35], Christine L. Clarke[70], NBCS Collaborators[41,71-84], Margriet Collée[85], Don M. Conroy[8], Emilie Cordina-Duverger[86], Sten Cornelissen[12], David G Cox[87,88], Angela Cox[50], Simon S. Cross[89], Julie M. Cunningham[90], Kamila Czene[42], Mary B. Daly[91], Peter Devilee[92,93], Kimberly F. Doheny[17], Thilo Dörk[36], Isabel dos-Santos-Silva[94], Martine Dumont[9], Lorraine Durcan[95,96], Miriam Dwek[97], Diana M. Eccles[96], Arif B. Ekici[98], A. Heather Eliassen [99,100], Carolina Ellberg[49,101], Mingajeva Elvira[102], Christoph Engel[103,104], Mikael Eriksson[42], Peter A. Fasching[29,105], Jonine Figueroa[48,106], Dieter Flesch-Janys[107,108], Olivia Fletcher[109], Henrik Flyger[110], Lin Fritschi[111], Valerie Gaborieau[46], Marike Gabrielson[42], Manuela Gago-Dominguez[60,112], Yu-Tang Gao[113], Susan M. Gapstur[62], José A. García-Sáenz[58], Mia M. Gaudet[62], Vassilios Georgoulias[114], Graham G. Giles[115,116], Gord Glendon[117], Mark S. Goldberg[118,119], David E. Goldgar[120], Anna González-Neira[30], Grethe I. Grenaker Alnæs[41], Mervi Grip[121], Jacek Gronwald[122], Anne Grundy[123], Pascal Guénel[86], Lothar Haeberle[29], Eric Hahnen[124-126], Christopher A. Haiman[127], Niclas Håkansson[128], Ute Hamann[129], Nathalie Hamel[28], Susan Hankinson[130], Patricia Harrington[8], Steven N. Hart[131], Jaana M. Hartikainen[132-134], Mikael Hartman[67,135], Alexander Hein[29], Jane Heyworth[136], Belynda Hicks[11], Peter Hillemanns[36], Dona N. Ho[65], Antoinette Hollestelle[137], Maartje J. Hooning[137], Robert N. Hoover[48], John L. Hopper[116], Ming-Feng Hou[138], Chia-Ni Hsiung[139], Guanmengqian Huang[129], Keith Humphreys[42], Junko Ishiguro[140,141], Hidemi Ito[140,141], Motoki Iwasaki[142], Hiroji Iwata[143], Anna Jakubowska[122], Wolfgang Janni[144], Esther M. John[145-147], Nichola Johnson[109], Kristine Jones[11], Michael Jones[148], Arja Jukkola-Vuorinen[149], Rudolf Kaaks[13], Maria Kabisch[129], Katarzyna Kaczmarek[122], Daehee Kang[68,69,150], Yoshio Kasuga[151], Michael J. Kerin[152], Sofia Khan[153], Elza Khusnutdinova[32,102], Johanna I. Kiiski[153], Sung-Won Kim[154], Julia A. Knight[155,156], Veli-Matti Kosma[132-134], Vessela N. Kristensen[41,77,78], Ute Krüger[49], Ava Kwong[64,157,158], Diether Lambrechts[159,160], Loic Le Marchand[161], Eunjung Lee[127], Min Hyuk Lee[162], Jong Won Lee[163], Chuen Neng Lee[135,164], Flavio Lejbkowicz[165], Jingmei Li[42], Jenna Lilyquist[131], Annika Lindblom[166], Jolanta Lissowska[167], Wing-Yee Lo[43,44], Sibylle Loibl[168], Jirong Long[57], Artitaya Lophatananon[169,170], Jan Lubinski[122], Craig Luccarini[8], Michael P. Lux[29], Edmond S.K. Ma[64,65], Robert J. MacInnis [115,116], Tom Maishman[95,96], Enes Makalic[116], Kathleen E Malone[171], Ivana Maleva Kostovska[172], Arto Mannermaa[132-134], Siranoush Manoukian[173], JoAnn E. Manson[100,174], Sara Margolin[175], Shivaani Mariapun[176], Maria Elena Martinez[112,177], Keitaro Matsuo[141,178], Dimitrios Mavroudis[114], James McKay[46], Catriona McLean[179], Hanne Meijers-Heijboer[18], Alfons Meindl[180], Primitiva Menéndez[181],

45 Usha Menon[182], Jeffery Meyer[90], Hui Miao[67], Nicola Miller[152], Nur Aishah Mohd Taib[183], Kenneth
46 Muir[169,170], Anna Marie Mulligan[184,185], Claire Mulot[186], Susan L. Neuhausen[33], Heli Nevanlinna[153],
47 Patrick Neven[187], Sune F. Nielsen[37,38], Dong-Young Noh[188], Børge G. Nordestgaard[37-39], Aaron
48 Norman[131], Olufunmilayo I. Olopade[189], Janet E. Olson[131], Håkan Olsson[49], Curtis Olswold[131], Nick
49 Orr[109], V. Shane Pankratz[190], Sue K. Park[68,69,150], Tjoung-Won Park-Simon[36], Rachel Lloyd[191], Jose I.A.
50 Perez[192], Paolo Peterlongo[193], Julian Peto[94], Kelly-Anne Phillips[116,194-196], Mila Pinchev[165], Dijana
51 Plaseska-Karanfilska[172], Ross Prentice[26], Nadege Presneau[97], Darya Prokofieva[102], Elizabeth Pugh[17],
52 Katri Pylkäs[197,198], Brigitte Rack[199], Paolo Radice[200], Nazneen Rahman[201], Gadi Rennert[165], Hedy S.
53 Rennert[165], Valerie Rhenius[8], Atocha Romero[58,202], Jane Romm[17], Kathryn J Ruddy[203], Thomas
54 Rüdiger[204], Anja Rudolph[13], Matthias Ruebner[29], Emiel J. Th. Rutgers[205], Emmanouil Saloustros[206],
55 Dale P. Sandler[207], Suleeporn Sangrajrang[208], Elinor J. Sawyer[209], Daniel F. Schmidt[116], Rita K.
56 Schmutzler[124-126], Andreas Schneeweiss[55,210], Minouk J. Schoemaker[148], Fredrick Schumacher[211], Peter
57 Schürmann[36], Rodney J. Scott[212,213], Christopher Scott[131], Sheila Seal[201], Caroline Seynaeve[137], Mitul
58 Shah[8], Priyanka Sharma[214], Chen-Yang Shen[215,216], Grace Sheng[127], Mark E. Sherman[217], Martha J.
59 Shrubsole[57], Xiao-Ou Shu[57], Ann Smeets[187], Christof Sohn[210], Melissa C. Southey[218], John J.
60 Spinelli[219,220], Christa Stegmaier[221], Sarah Stewart-Brown[169], Jennifer Stone[191,222], Daniel O. Stram[127],
61 Harald Surowy[55,56], Anthony Swerdlow[148,223], Rulla Tamimi[5,99,100], Jack A. Taylor[207,224], Maria
62 Tengström[132,225,226], Soo H. Teo[176,183], Mary Beth Terry[227], Daniel C. Tessier[28], Somchai
63 Thanasitthichai[228], Kathrin Thöne[108], Rob A.E.M. Tollenaar[229], Ian Tomlinson[230], Ling Tong[19], Diana
64 Torres[129,231], Thérèse Truong[86], Chiu-chen Tseng[127], Shoichiro Tsugane[232], Hans-Ulrich Ulmer[233], Giske
65 Ursin[234,235], Michael Untch[236], Celine Vachon[131], Christi J. van Asperen[237], David Van Den Berg[127], Ans
66 M.W. van den Ouweland[85], Lizet van der Kolk[238], Rob B. van der Luijt[239], Daniel Vincent[28], Jason
67 Vollenweider[90], Quinten Waisfisz[18], Shan Wang-Gohrke[240], Clarice R. Weinberg[241], Camilla Wendt[175],
68 Alice S. Whittemore[146,147], Hans Wildiers[187], Walter Willett [100,242], Robert Winqvist[197,198], Alicja
69 Wolk[128], Anna H. Wu[127], Lucy Xia[127], Taiki Yamaji[142], Xiaohong R. Yang[48], Cheng Har Yip[243], Keun-Young
70 Yoo[244,245], Jyh-Cherng Yu[246], Wei Zheng[57], Ying Zheng[247], Bin Zhu[11], Argyrios Ziogas[21], Elad Ziv[248],
71 ABCTB Investigators [249], kConFab/AOCS Investigators [194,250], Sunil R. Lakhani[14,251], Antonis C.
72 Antoniou[1], Arnaud Droit[9], Irene L. Andrulis[117,252], Christopher I. Amos[253], Fergus J. Couch[90], Paul D.P.
73 Pharoah[1,8], Jenny Chang-Claude[13,254], Per Hall[42,255], David J. Hunter[5,100], Roger L. Milne[115,116],
74 Montserrat García-Closas[48], Marjanka K. Schmidt[12,256], Stephen J. Chanock[48], Alison M. Dunning[8],
75 Stacey L. Edwards[6], Gary D. Bader[7], Georgia Chenevix-Trench[6], Jacques Simard[9,257], Peter Kraft[5,100,257],
76 Douglas F. Easton[1,8,257].

77

78 1. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care,
79 University of Cambridge, Cambridge, UK.
80 2. Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology
81 and Genetics, Nicosia, Cyprus.
82 3. These authors contributed equally to this work.
83 4. Department of Epidemiology, University of Washington School of Public Health, Seattle, WA,
84 USA.
85 5. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public
86 Health, Boston, MA, USA.
87 6. Department of Genetics and Computational Biology, QIMR Berghofer Medical Research
88 Institute, Brisbane, Australia.
89 7. The Donnelly Centre, University of Toronto, Toronto, ON, Canada.

90 8. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge,
91 Cambridge, UK.
92 9. Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval
93 University, Québec City, QC, Canada.
94 10. Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN,
95 USA.
96 11. Cancer Genomics Research Laboratory (CGR), Division of Cancer Epidemiology and Genetics,
97 National Cancer Institute, Rockville, MD, USA.
98 12. Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van
99 Leeuwenhoek Hospital, Amsterdam, The Netherlands.
100 13. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg,
101 Germany.
102 14. UQ Centre for Clinical Research, The University of Queensland, Brisbane, Australia.
103 15. Cancer Research UK Cambridge Research Institute, University of Cambridge, Li Ka Shing
104 Centre, Cambridge, UK.
105 16. Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA.
106 17. Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins
107 University School of Medicine, Baltimore, MD, USA.
108 18. Department of Clinical Genetics, VU University Medical Center, Amsterdam, The
109 Netherlands.
110 19. Center for Cancer Epidemiology and Prevention, The University of Chicago, Chicago, IL, USA.
111 20. Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki,
112 Finland.
113 21. Department of Epidemiology, University of California Irvine, Irvine, CA, USA.
114 22. N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus.
115 23. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center
116 (DKFZ), Heidelberg, Germany.
117 24. Department of Public Health Sciences, and Cancer Research Institute, Queen's University,
118 Kingston, ON, Canada.
119 25. Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center,
120 Houston, TX, USA.
121 26. Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.
122 27. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA.
123 28. McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada.
124 29. Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-
125 Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN,
126 Erlangen, Germany.
127 30. Human Cancer Genetics Program, Spanish National Cancer Research Centre, Madrid, Spain.
128 31. Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain.
129 32. Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences,
130 Ufa, Russia.
131 33. Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA,
132 USA.
133 34. Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki,
134 Finland.
135 35. Department of Radiation Oncology, Hannover Medical School, Hannover, Germany.
136 36. Gynaecology Research Unit, Hannover Medical School, Hannover, Germany.
137 37. Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen
138 University Hospital, Herlev, Denmark.
139 38. Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University
140 Hospital, Herlev, Denmark.

141 39. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
142 40. Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, Milan, Italy.
143 41. Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital
144 Radiumhospitalet, Oslo, Norway.
145 42. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm,
146 Sweden.
147 43. Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany.
148 44. University of Tübingen, Tübingen, Germany.
149 45. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg,
150 Germany.
151 46. International Agency for Research on Cancer, Lyon, France.
152 47. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National
153 Center for Tumor Diseases (NCT), Heidelberg, Germany.
154 48. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA.
155 49. Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden.
156 50. Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism,
157 University of Sheffield, Sheffield, UK.
158 51. Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada.
159 52. Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC,
160 Canada.
161 53. Department of Gynecology and Obstetrics, University of Tübingen, Tübingen, Germany.
162 54. Institute for Prevention and Occupational Medicine of the German Social Accident
163 Insurance, Institute of the Ruhr University Bochum, Bochum, Germany.
164 55. Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, Germany.
165 56. Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg,
166 Germany.
167 57. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center,
168 Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN,
169 USA.
170 58. Medical Oncology Department, Hospital Clínico San Carlos, IdISSC (Centro Investigacion
171 Biomedica en Red), CIBERONC (Instituto de Investigación Sanitaria San Carlos), Madrid,
172 Spain.
173 59. Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg,
174 Germany.
175 60. Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de
176 Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario
177 Universitario de Santiago, SERGAS, Santiago De Compostela, Spain.
178 61. Centro de Investigación en Red de Enfermedades Raras (CIBERER) y Centro Nacional de
179 Genotipado (CEGEN-PRB2), Universidad de Santiago de Compostela, Santiago De
180 Compostela, Spain.
181 62. Epidemiology Research Program, American Cancer Society, Atlanta, GA, USA.
182 63. Oncology and Genetics Unit, Instituto de Investigacion Biomedica (IBI) Orense-Pontevedra-
183 Vigo, Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain.
184 64. Hong Kong Hereditary Breast Cancer Family Registry, Happy Valley, Hong Kong.
185 65. Department of Pathology, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong.
186 66. Division of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA.
187 67. Saw Swee Hock School of Public Health, National University of Singapore, Singapore,
188 Singapore.
189 68. Department of Biomedical Sciences, Seoul National University Graduate School, Seoul,
190 Korea.
191 69. Cancer Research Institute, Seoul National University, Seoul, Korea.

192 70. Westmead Institute for Medical Research, University of Sydney, Sydney, Australia.
193 71. Department of Oncology, Haukeland University Hospital, Bergen, Norway.
194 72. Section of Oncology, Institute of Medicine, University of Bergen, Bergen, Norway.
195 73. Department of Pathology, Akershus University Hospital, Lørenskog, Norway.
196 74. Department of Breast-Endocrine Surgery, Akershus University Hospital, Lørenskog, Norway.
197 75. Department of Breast and Endocrine Surgery, Oslo University Hospital, Ullevål, Oslo,
198 Norway.
199 76. Department of Research, Vestre Viken, Drammen, Norway.
200 77. Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway.
201 78. Department of Clinical Molecular Biology, Oslo University Hospital, University of Oslo, Oslo,
202 Norway.
203 79. National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital
204 Radiumhospitalet, Oslo, Norway.
205 80. Department of Oncology, Oslo University Hospital Radiumhospitalet, Oslo, Norway.
206 81. Department of Radiology and Nuclear Medicine, Oslo University Hospital Radiumhospitalet,
207 Oslo, Norway.
208 82. Oslo University Hospital, Oslo, Norway.
209 83. Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital
210 Radiumhospitalet, Oslo, Norway.
211 84. Department of Oncology, Oslo University Hospital Ullevål, Oslo, Norway.
212 85. Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The
213 Netherlands.
214 86. Cancer & Environment Group, Center for Research in Epidemiology and Population Health
215 (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France.
216 87. Department of Epidemiology and Biostatistics, School of Public Health, Imperial College
217 London, London, UK.
218 88. INSERM U1052, Cancer Research Center of Lyon, Lyon, France.
219 89. Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield,
220 UK.
221 90. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA.
222 91. Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA.
223 92. Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.
224 93. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.
225 94. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and
226 Tropical Medicine, London, UK.
227 95. Southampton Clinical Trials Unit, Faculty of Medicine , University of Southampton,
228 Southampton, UK.
229 96. Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton,
230 Southampton, UK.
231 97. Department of Biomedical Sciences, Faculty of Science and Technology, University of
232 Westminster, London, UK.
233 98. Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University
234 Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany.
235 99. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's
236 Hospital, Harvard Medical School, Boston, MA, USA.
237 100. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
238 101. Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden.
239 102. Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia.
240 103. Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig,
241 Germany.

242 104. LIFE - Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig,
243 Germany.
244 105. David Geffen School of Medicine, Department of Medicine Division of Hematology and
245 Oncology, University of California at Los Angeles, Los Angeles, CA, USA.
246 106. Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh
247 Medical School, Edinburgh, UK.
248 107. Institute for Medical Biometrics and Epidemiology, University Medical Center Hamburg-
249 Eppendorf, Hamburg, Germany.
250 108. Department of Cancer Epidemiology, Clinical Cancer Registry, University Medical Center
251 Hamburg-Eppendorf, Hamburg, Germany.
252 109. Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London,
253 UK.
254 110. Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University
255 Hospital, Herlev, Denmark.
256 111. School of Public Health, Curtin University, Perth, Australia.
257 112. Moores Cancer Center, University of California San Diego, La Jolla, CA, USA.
258 113. Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China.
259 114. Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece.
260 115. Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, Melbourne, Victoria,
261 Australia.
262 116. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global
263 Health, The University of Melbourne, Melbourne, Australia.
264 117. Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of
265 Mount Sinai Hospital, Toronto, ON, Canada.
266 118. Department of Medicine, McGill University, Montréal, QC, Canada.
267 119. Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, QC,
268 Canada.
269 120. Department of Dermatology, Huntsman Cancer Institute, University of Utah School of
270 Medicine, Salt Lake City, UT, USA.
271 121. Department of Surgery, Oulu University Hospital, University of Oulu, Oulu, Finland.
272 122. Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland.
273 123. Centre de Recherche du Centre Hospitalier de Université de Montréal (CHUM), Université de
274 Montréal, Montréal, QC, Canada.
275 124. Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne, Cologne,
276 Germany.
277 125. Center for Integrated Oncology (CIO), University Hospital of Cologne, Cologne, Germany.
278 126. Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany.
279 127. Department of Preventive Medicine, Keck School of Medicine, University of Southern
280 California, Los Angeles, CA, USA.
281 128. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
282 129. Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg,
283 Germany.
284 130. Department of Biostatistics & Epidemiology, University of Massachusetts, Amherst, Amherst,
285 MA, USA.
286 131. Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.
287 132. Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland.
288 133. Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern
289 Finland, Kuopio, Finland.
290 134. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio,
291 Finland.
292 135. Department of Surgery, National University Health System, Singapore, Singapore.

293 136. School of Population Health, University of Western Australia, Perth, Australia.
294 137. Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute,
295 Rotterdam, The Netherlands.
296 138. Department of Surgery, Kaohsiung Municipal Hsiao-Kang Hospital, Kaohsiung, Taiwan.
297 139. Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan.
298 140. Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya,
299 Japan.
300 141. Department of Epidemiology, Nagoya University Graduate School of Medicine, Nagoya,
301 Japan.
302 142. Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo,
303 Japan.
304 143. Department of Breast Oncology, Aichi Cancer Center Hospital, Nagoya, Japan.
305 144. Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm, Germany.
306 145. Department of Epidemiology, Cancer Prevention Institute of California, Fremont, CA, USA.
307 146. Department of Health Research and Policy - Epidemiology, Stanford University School of
308 Medicine, Stanford, CA, USA.
309 147. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA.
310 148. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK.
311 149. Department of Oncology, Oulu University Hospital, University of Oulu, Oulu, Finland.
312 150. Department of Preventive Medicine, Seoul National University College of Medicine, Seoul,
313 Korea.
314 151. Department of Surgery, Nagano Matsushiro General Hospital, Nagano, Japan.
315 152. School of Medicine, National University of Ireland, Galway, Ireland.
316 153. Department of Obstetrics and Gynecology, Helsinki University Hospital, University of
317 Helsinki, Helsinki, Finland.
318 154. Department of Surgery, Daerim Saint Mary's Hospital, Seoul, Korea.
319 155. Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute of Mount
320 Sinai Hospital, Toronto, ON, Canada.
321 156. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto,
322 ON, Canada.
323 157. Department of Surgery, The University of Hong Kong, Pok Fu Lam, Hong Kong.
324 158. Department of Surgery, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong.
325 159. Vesalius Research Center, VIB, Leuven, Belgium.
326 160. Laboratory for Translational Genetics, Department of Oncology, University of Leuven,
327 Leuven, Belgium.
328 161. University of Hawaii Cancer Center, Honolulu, HI, USA.
329 162. Department of Surgery, Soonchunhyang University College of Medicine and Soonchunhyang
330 University Hospital, Seoul, Korea.
331 163. Department of Surgery, Ulsan University College of Medicine and Asan Medical Center,
332 Seoul, Korea.
333 164. Department of Cardiac, Thoracic and Vascular Surgery, National University Health System,
334 Singapore, Singapore.
335 165. Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of
336 Medicine, Haifa, Israel.
337 166. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden.
338 167. Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Memorial Cancer
339 Center & Institute of Oncology, Warsaw, Poland.
340 168. German Breast Group, GmbH, Neu Isenburg, Germany.
341 169. Division of Health Sciences, Warwick Medical School, Warwick University, Coventry, UK.
342 170. Institute of Population Health, University of Manchester, Manchester, UK.

343 171. Division of Public Health Sciences, Epidemiology Program, Fred Hutchinson Cancer Research
344 Center, Seattle, WA, USA.
345 172. Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov" ,
346 Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia.
347 173. Unit of Medical Genetics, Department of Preventive and Predictive Medicine, Fondazione
348 IRCCS (Istituto Di Ricovero e Cura a Carattere Scientifico) Istituto Nazionale dei Tumori (INT),
349 Milan, Italy.
350 174. Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston,
351 MA, USA.
352 175. Department of Oncology - Pathology, Karolinska Institutet, Stockholm, Sweden.
353 176. Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia.
354 177. Department of Family Medicine and Public Health, University of California San Diego, La
355 Jolla, CA, USA.
356 178. Division of Molecular Medicine, Aichi Cancer Center Research Institute, Nagoya, Japan.
357 179. Anatomical Pathology, The Alfred Hospital, Melbourne, Australia.
358 180. Division of Gynaecology and Obstetrics, Technische Universität München, Munich, Germany.
359 181. Servicio de Anatomía Patológica, Hospital Monte Naranco, Oviedo, Spain.
360 182. Gynaecological Cancer Research Centre, Department for Women's Cancer, Institute for
361 Women's Health, University College London, London, UK.
362 183. Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre,
363 Kuala Lumpur, Malaysia.
364 184. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON,
365 Canada.
366 185. Laboratory Medicine Program, University Health Network, Toronto, ON, Canada.
367 186. Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris, France.
368 187. Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute,
369 University Hospitals Leuven, Leuven, Belgium.
370 188. Department of Surgery, Seoul National University College of Medicine, Seoul, Korea.
371 189. Center for Clinical Cancer Genetics and Global Health, The University of Chicago, Chicago, IL,
372 USA.
373 190. University of New Mexico Health Sciences Center, University of New Mexico, Albuquerque,
374 NM, USA.
375 191. The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and
376 University of Western Australia, Perth, Australia.
377 192. Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, Oviedo, Spain.
378 193. IFOM, The FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology,
379 Milan, Italy.
380 194. Peter MacCallum Cancer Center, Melbourne, Australia.
381 195. Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne,
382 Australia.
383 196. Department of Medicine, St Vincent's Hospital, The University of Melbourne, Fitzroy,
384 Australia.
385 197. Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine
386 Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland.
387 198. Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu,
388 Oulu, Finland.
389 199. Department of Gynecology and Obstetrics, Ludwig-Maximilians University of Munich,
390 Munich, Germany.
391 200. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and
392 Predictive Medicine, Fondazione IRCCS (Istituto Di Ricovero e Cura a Carattere Scientifico)
393 Istituto Nazionale dei Tumori (INT), Milan, Italy.

394 201. Section of Cancer Genetics, The Institute of Cancer Research, London, UK.
395 202. Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain.
396 203. Department of Oncology, Mayo Clinic, Rochester, MN, USA.
397 204. Institute of Pathology, Staedtisches Klinikum Karlsruhe, Karlsruhe, Germany.
398 205. Department of Surgery, The Netherlands Cancer Institute - Antoni van Leeuwenhoek
399 hospital, Amsterdam, The Netherlands.
400 206. Hereditary Cancer Clinic, University Hospital of Heraklion, Heraklion, Greece.
401 207. Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research
402 Triangle Park, NC, USA.
403 208. National Cancer Institute, Bangkok, Thailand.
404 209. Research Oncology, Guy's Hospital, King's College London, London, UK.
405 210. National Center for Tumor Diseases, University of Heidelberg, Heidelberg, Germany.
406 211. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland,
407 OH, USA.
408 212. Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle,
409 Australia.
410 213. Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of
411 Health, University of Newcastle, Callaghan, Australia.
412 214. Department of Medicine, Kansas University Medicial Center, Kansas City, KS, USA.
413 215. School of Public Health, China Medical University, Taichung, Taiwan.
414 216. Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan.
415 217. Division of Cancer Prevention, National Cancer Institute, Rockville, MD, USA.
416 218. Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne,
417 Melbourne, Australia.
418 219. Cancer Control Research, BC Cancer Agency, Vancouver, BC, Canada.
419 220. School of Population and Public Health, University of British Columbia, Vancouver, BC,
420 Canada.
421 221. Saarland Cancer Registry, Saarbrücken, Germany.
422 222. Department of Obstetrics and Gynaecology, University of Melbourne and the Royal
423 Women's Hospital, Melbourne, Australia.
424 223. Division of Breast Cancer Research, The Institute of Cancer Research, London, UK.
425 224. Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health
426 Sciences, NIH, Research Triangle Park, NC, USA.
427 225. Cancer Center, Kuopio University Hospital, Kuopio, Finland.
428 226. Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio, Finland.
429 227. Department of Epidemiology, Mailman School of Public Health, Columbia University, New
430 York, NY, USA.
431 228. National Cancer Institute, Ministry of Public Health, Nonthaburi, Thailand.
432 229. Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands.
433 230. Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre,
434 University of Oxford, Oxford, UK.
435 231. Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia.
436 232. Center for Public Health Sciences, National Cancer Center, Tokyo, Japan.
437 233. Frauenklinik der Stadtklinik Baden-Baden, Baden-Baden, Germany.
438 234. Cancer Registry of Norway, Oslo, Norway
439 235. Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo,
440 Norway.
441 236. Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany.
442 237. Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands.
443 238. Family Cancer Clinic, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital,
444 Amsterdam, The Netherlands.

445 239. Division of Biomedical Genetics, University Medical Center Utrecht, Utrecht, The
446 Netherlands.
447 240. Department of Gynaecology and Obstetrics, University of Ulm, Ulm, Germany.
448 241. Biostatistics and Computational Biology Branch, National Institute of Environmental Health
449 Sciences, NIH, Research Triangle Park, NC, USA.
450 242. Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
451 243. Subang Jaya Medical Centre, Subang Jaya, Selangor, Malaysia.
452 244. Seoul National University College of Medicine, Seoul, Korea.
453 245. Armed Forces Capital Hospital, Seongnam, Korea.
454 246. Department of Surgery, Tri-Service General Hospital, National Defense Medical Center,
455 Taipei, Taiwan.
456 247. Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China.
457 248. Department of Medicine, Institute for Human Genetics, UCSF Helen Diller Family
458 Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA,
459 USA.
460 249. Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research, University
461 of Sydney, Sydney, Australia.
462 250. QIMR Berghofer Medical Research Institute, Brisbane, Australia.
463 251. Pathology Queensland, The Royal Brisbane and Women's Hospital, Brisbane 4029, Australia.
464 252. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada.
465 253. Center for Genomic Medicine, Department of Biomedical Data Science, Geisel School of
466 Medicine, Dartmouth College, Hanover, NH, USA.
467 254. University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf,
468 Hamburg, Germany.
469 255. Department of Oncology, Södersjukhuset, Stockholm, Sweden.
470 256. Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute -
471 Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands.
472 257. These authors jointly supervised this work.

473

474

475 Correspondence should be addressed to D.F.E. (dfe20@medschl.cam.ac.uk)

476

Breast cancer risk is influenced by rare coding variants in susceptibility genes such as *BRCA1* and many common, mainly non-coding variants. However, much of the genetic contribution to breast cancer risk remains unknown. We report results from a genome-wide association study (GWAS) of breast cancer in 122,977 cases and 105,974 controls of European ancestry and 14,068 cases and 13,104 controls of East Asian ancestry[1]. We identified 65 new loci associated with overall breast cancer at $p<5x10^{-8}$. The majority of credible risk SNPs in the new loci fall in distal regulatory elements, and by integrating *in-silico* data to predict target genes in breast cells at each locus, we demonstrate a strong overlap between candidate target genes and somatic driver genes in breast tumours. We also find that heritability of breast cancer due to all SNPs in regulatory features was 2-5-fold enriched relative to the genome-wide average, with strong enrichment for particular transcription factor binding sites. These results provide further insight into genetic susceptibility to breast cancer and will improve the utility of genetic risk scores for individualized screening and prevention.

492 We genotyped 61,282 female breast cancer cases and 45,494 female controls of European ancestry
493 with the OncoArray[1]. Subjects came from 68 studies collaborating in the Breast Cancer Association
494 Consortium (BCAC) and Discovery, Biology and Risk of Inherited Variants in Breast Cancer
495 Consortium (DRIVE) **(Supplementary Table 1)**. Using the 1000 Genomes Project (Phase 3) reference
496 panel, we imputed genotypes for ~21M variants. After filtering on minor allele frequency
497 (MAF)>0.5% and imputation quality score>0.3 (see **Online Methods**), we assessed the association
498 between breast cancer risk and 11.8M SNPs adjusting for country and ancestry-informative principal
499 components. We combined these results with results from the iCOGS project (46,785 cases and
500 42,892 controls)[2] and 11 other breast cancer GWAS (14,910 cases, 17,588 controls), using a fixed-
501 effect meta-analysis.
502
503 Of 102 loci previously associated with breast cancer in Europeans, 49 showed evidence for
504 association with overall breast cancer in the OncoArray dataset at $P<5\times10^{-8}$ and 94 at $P<0.05$. Five
505 additional loci previously shown to be associated with breast cancer in Asian women also showed
506 evidence in the European ancestry OncoArray dataset ($P<0.01$; **Supplementary Tables 2-4**)[3-5]. We
507 also assessed the association with breast cancer in Asians including 7,799 cases and 6,480 controls
508 from the OncoArray project and 6,269 cases and 6,624 controls from iCOGS. Of the 94 loci previously
509 identified in Europeans that were polymorphic in Asians, 50 showed evidence of association
510 ($P<0.05$). For the remaining 44, none showed a significant difference in the estimated odds ratio
511 (OR) for overall breast cancer between Europeans and Asians ($P>0.01$; **Supplementary Table 5**). The
512 correlation in effect sizes for all known loci between Europeans and Asians was 0.83, suggesting that
513 the majority of known susceptibility loci are shared between these populations.
514
515 To search for additional susceptibility loci, we assessed all SNPs excluding those within 500kb of a
516 known susceptibility SNPs (**Figure 1**). This identified 5,969 variants in 65 regions that were associated
517 with overall breast cancer risk at $P<5\times10^{-8}$ (**Table 1, Supplementary Tables 6-7**). For two loci (lead
518 SNPs rs58847541 and rs12628403), there was evidence of a second association signal after
519 adjustment for the primary signal (rs13279803: conditional $P=1.6\times10^{-10}$; rs373038216: $P=2.9\times10^{-11}$;
520 **Supplementary Table 8**). Of the 65 new loci, 21 showed a differential association by ER-status
521 ($P<0.05$) with all but two (rs6725517 and rs6569648) more strongly associated with ER-positive
522 disease (**Supplementary Tables 9-10**). Forty-four loci showed evidence of association for ER-negative
523 breast cancer ($P<0.05$). Of the 51 novel loci that were polymorphic in Asians, nine were associated at
524 $P<0.05$ and only two showed a difference in the estimated OR between Europeans and Asians
525 ($P<0.01$; **Supplementary Table 11**).

526 To define a set of credible risk variants (CRVs) at the new loci, we first selected variants with *P*-values
527 within two orders of magnitude of the most significant SNPs in each region. Across the 65 novel
528 regions, we identified 2,221 CRVs (**Supplementary Table 12**), while the previous 77 identified loci
529 contained 2,232 CRVs (**Online methods; Supplementary Table 13**). We examined the evidence for
530 enrichment in these CRVs of 67 genomic features, including histone marks and transcription factor
531 binding sites (TFBS) in three breast cancer cell lines (**Online Methods; Supplementary Tables 14-15;**
532 **Extended Data Fig. 1**). Thirteen features were significant predictors of CRVs at $P<10^{-4}$; the strongest
533 being DNAse I hypersensitivity sites in CTCF silenced MCF7 cells (OR 2.38, $P=4.6\times10^{-14}$). Strong
534 associations were also observed with binding sites for FOXA1, ESR1, GATA3, E2F1 and TCF7L2. Seven
535 of the 65 novel loci included only a single CRV (**Supplementary Table 6**), of which two are non-
536 synonymous. SNP rs16991615 is a missense variant (p.Glu341Lys) in *MCM8*, involved in genome
537 replication and associated with age at natural menopause and impaired DNA repair[6]. SNP
538 rs35383942 is a missense variant (p.Arg28Gln) in *PHLDA3*, encoding a p53-regulated repressor of
539 AKT[7].
540
541 We annotated each CRV with publicly available genomic data from breast cells in order to highlight
542 potentially functional variants, predict target genes and prioritise future experimental validation
543 (**Supplementary Tables 6 and 12** with UCSC browser links). We developed a heuristic scoring system
544 based on breast-specific genomic data (integrated expression quantitative trait and *in silico*
545 prediction of GWAS targets **-** INQUISIT) to rank the target genes at each locus (**Supplementary Table**
546 **16)**. Target genes were predicted by combining risk SNP data with multiple sources of genomic
547 information, including chromatin interactions (ChIA-PET and Hi-C), computational enhancer-
548 promoter correlations (PreSTIGE, IM-PET, FANTOM5 and Super-enhancers), breast tissue-specific
549 eQTL results, TF binding (ENCODE ChIP-seq), gene expression (ENCODE RNA-seq) and topologically-
550 associated domain (TAD) boundaries (**Online Methods** and **Supplementary Tables 17-19**). Target
551 gene predictions could be made for 58/65 new and 70/77 previously identified loci. Among 689
552 protein-coding genes predicted by INQUISIT, we found strong enrichment for established breast
553 cancer drivers identified through tumour sequencing (20/147 genes, $P<10^{-6}$)[8-11], which increased
554 with increasing INQUISIT score ($P=1.8\times10^{-6}$). We compared INQUISIT with a) an alternative published
555 method (DEPICT, which predicts targets based on shared gene functions between potential targets
556 at other associated loci)[12] which showed a weaker enrichment of breast cancer driver genes ($P=0.06$
557 after adjusting for the nearest gene, $P=0.74$ after adjusting for INQUIST score, and b) assigning the
558 association signal to the nearest gene, which showed only a weak enrichment of driver genes after
559 adjusting for the INQUISIT score ($P=0.01$; **Extended Data Table 1** and **Supplementary Table 20**).

560 Notably, most of the 689 putative target genes have no reported involvement in breast
561 tumorigenesis and some may represent additional genes influencing susceptibility to breast cancer.
562 However, functional assays will be required to confirm any of these candidates as risk genes.
563
564 Having used INQUISIT to predict target genes, we performed pathway gene set enrichment analysis
565 (GSEA), visually summarized as enrichment maps (**Extended Data Fig. 2; Supplementary Tables 21-**
566 **22**)[13]. Several growth or development related pathways were enriched, notably the fibroblast
567 growth factor, platelet derived growth factor and Wnt signalling pathways[14-16]. Other cancer-related
568 themes included ERK1/2 cascade, immune-response pathways including interferon signalling, and
569 cell-cycle pathways. Pathways not found in earlier breast cancer GWAS include nitric oxide
570 biosynthesis, AP-1 transcription factor and NF-kB (**Supplementary Table 23**).
571
572 To explore more globally the genomic features contributing to breast cancer risk, we estimated the
573 proportion of genome-wide SNP heritability attributable to 53 publicly available annotations[17]. We
574 observed the largest enrichment in heritability (5.2-fold, P=8.5x10^{-5}) for TFBS, followed by a 4-fold
575 (P=0.0006) enrichment for histone marker H3K4me3 (marking promoters). In contrast, we observed
576 a significant depletion (0.27, P=0.0007) for repressed regions (**Supplementary Table 24).** We
577 conducted cell type-specific enrichment analysis for four histone marks and observed significant
578 enrichments in several tissue types (**Figure 2; Extended Data Figs. 3-7; Supplementary Table 25-26**),
579 including a 6.7-fold enrichment for H3K4me1 in breast myoepithelial tissue (P=7.9x10^{-5}). We
580 compared the cell type-specific enrichments for overall, ER-positive and ER-negative breast cancer to
581 the enrichments for 16 other complex traits (**Extended Data Figs. 3-7**). Breast cancer showed
582 enrichment for adipose and epithelial cell types (including breast epithelial cells). In contrast,
583 psychiatric diseases showed enrichment specific to central-nervous-system cell types and
584 autoimmune disorders showed enrichment for immune cells.
585
586 We selected for further evaluation four loci to represent those predicted to act through proximal
587 regulation (1p36 and 11p15) and distal regulation (1p34 and 7q22), because they had a relatively
588 small number of CRVs. The only CRV at 1p36, rs2992756 (P=1.6x10^{-15}), is located 84bp from the
589 transcription start site of *KLHDC7A*. Of the 19 CRVs at 11p15 (smallest P=1.4x10^{-12}), five were located
590 in the proximal promoter of *PIDD1*, implicated in DNA-damage-induced apoptosis and
591 tumorigenesis[18]. INQUIST predicted *KLHDC7A* and *PIDD1* to be target genes and they received the
592 highest score for likelihood of promoter regulation (**Supplementary Table 18**). Using reporter assays,
593 we showed that the *KLHDC7A* promoter construct containing the risk *T*-allele of rs2992756 has

594 significantly lower activity than the reference construct, while the *PIDD1* promoter construct
595 containing the risk haplotype significantly increased *PIDD1* promoter activity (**Extended Data Fig. 8**).
596
597 The 1p34 locus included four CRVs (smallest P=9.1x10^{-9}) that fall within two putative regulatory
598 elements (PREs) and are predicted by INQUISIT to regulate *CITED4* (PREs; **Extended Data Fig. 8**).
599 CITED4 encodes a transcriptional coactivator that interacts with CBP/p300 and TFAP2 and can inhibit
600 hypoxia-activated transcription in cancer cells[19]. Chromatin conformation capture (3C) assays
601 confirmed that the PREs physically interacted with the *CITED4* promoter (**Extended Data Fig. 8)**.
602 Subsequent reporter assays showed that the PRE1 reference construct reduced *CITED4* promoter
603 activity, whereas the risk *T*-allele of SNP rs4233486 located in PRE1 negates this effect.
604
605 Finally, the 7q22 risk locus contained six CRVs (smallest P=5.1x10^{-12}) which lie in several PREs
606 spanning ~40kb of *CUX1* intron 1. Chromatin interactions were identified between a PRE1
607 (containing SNP rs6979850) and *CUX1*/*RASA4* promoters and a PRE2 (containing SNP rs71559437)
608 and *RASA4/PRKRIP1* promoters (**Extended Data Fig. 9**). Allele-specific 3C in heterozygous MBA-MB-
609 231 cells showed that the risk haplotype was associated with chromatin looping, suggesting that the
610 protective allele abrogates looping between the PREs and target genes (**Extended Data Fig. 9).** These
611 results identify two mechanisms by which CRVs may impact target gene expression: through
612 transactivation of a specific promoter and by affecting chromatin looping between regulatory
613 elements and their target genes. These data provide *in vitro* evidence of target identification and
614 regulation, however further studies that include genome editing, oncogenic assays and/or animal
615 models will be required to fully elucidate disease-related gene function.
616
617 We estimate that the newly identified susceptibility loci explain ~4% of the two-fold familial relative
618 risk (FRR) of breast cancer and that in total, common susceptibility variants identified through GWAS
619 explain 18% of the FRR. Further, we estimate that variants imputable from the OncoArray, under a
620 log-additive model (see **Online Methods**), explain ~41% of the FRR, and thus, the identified
621 susceptibility SNPs account for ~44% (18%/41%) of the FRR that can be explained by all imputable
622 SNPs. The identified SNPs will be incorporated into risk prediction models, which can be used to
623 improve the identification of women at high and low risk of breast cancer: for example, using a
624 polygenic risk score based on the variants identified to date, women in the highest 1% of the
625 distribution have a 3.5-fold greater breast cancer risk than the population average. Such risk
626 prediction can inform targeted early detection and prevention.

627 **Table 1.** Newly identified susceptibility loci for overall breast cancer[1].

Locus	Variant[1]	Chr[2]	Position[3]	Alleles[4]	MAF[5]	GWAS		iCOGS		OncoArray		Combined *P*-value	Genes[8]
						OR (95%CI)[6]	***P*[7]**	**OR (95%CI)[6]**	***P*[7]**	**OR (95%CI)[6]**	***P*[7]**		
1p36.13	rs2992756	1	18807339	C/T	0.49	1.03(0.99-1.06)	1.4x10^{-01}	1.05(1.03-1.07)	1.3x10^{-05}	1.06(1.04-1.08)	1.3x10^{-11}	1.6x10^{-15}	*KLHDC7A*
1p34.2	rs4233486	1	41380440	T/C	0.36	0.97(0.93-1)	6.6x10^{-02}	0.95(0.93-0.97)	3.6x10^{-05}	0.97(0.95-0.98)	2.3x10^{-04}	9.1x10^{-09}	-
1p34.2	rs79724016	1	42137311	T/G	0.03	0.85(0.77-0.95)	3.3x10^{-03}	0.90(0.85-0.95)	1.1x10^{-04}	0.93(0.88-0.97)	3.3x10^{-03}	3.5x10^{-08}	*HIVEP3*
1p34.1	rs1707302	1	46600917	G/A	0.34	0.97(0.93-1)	7.2x10^{-02}	0.96(0.94-0.98)	3.1x10^{-04}	0.96(0.95-0.98)	1.4x10^{-04}	3.0x10^{-08}	*PIK3R3, LOC101929626*
1p32.3	rs140850326	1	50846032	I/D[9]	0.49	0.94(0.91-0.98)	1.5x10^{-03}	0.97(0.95-0.99)	2.3x10^{-03}	0.97(0.95-0.99)	3.4x10^{-04}	3.9x10^{-08}	-
1p22.3	rs17426269	1	88156923	G/A	0.15	1.06(1.01-1.12)	1.1x10^{-02}	1.05(1.02-1.08)	6.6x10^{-04}	1.05(1.02-1.07)	1.7x10^{-04}	1.7x10^{-08}	-
1p12	rs7529522	1	118230221	T/C	0.23	1.06(1.01-1.12)	1.4x10^{-02}	1.03(1.01-1.05)	8.7x10^{-03}	1.06(1.04-1.08)	1.6x10^{-08}	1.7x10^{-10}	-
1q22	rs4971059	1	155148781	G/A	0.35	1.07(1.03-1.11)	3.7x10^{-04}	1.02(1-1.05)	1.4x10^{-02}	1.05(1.03-1.07)	3.9x10^{-08}	4.8x10^{-11}	*TRIM46*
1q32.1	rs35383942	1	201437832	C/T	0.06	1.08(0.99-1.17)	7.0x10^{-02}	1.09(1.04-1.14)	1.9x10^{-04}	1.12(1.08-1.17)	12x10^{-09}	3.8x10^{-13}	*PHLDA3*
1q41	rs11117758	1	217220574	G/A	0.21	0.95(0.91-0.99)	2.3x10^{-02}	0.97(0.95-0.99)	7.8x10^{-03}	0.95(0.93-0.97)	7.7x10^{-07}	3.9x10^{-09}	*ESRRG*
2p25.1	rs113577745	2	10135681	C/G	0.1	1.08(1.02-1.14)	8.9x10^{-03}	1.05(1.02-1.08)	3.7x10^{-03}	1.08(1.05-1.11)	3.7x10^{-07}	3.9x10^{-10}	*GRHL1*
2p23.3	rs6725517	2	25129473	A/G	0.41	0.95(0.91-0.98)	1.8x10^{-3}	0.95(0.93-0.97)	8.5x10^{-06}	0.96(0.94-0.98)	7.5x10^{-06}	2.9x10^{-12}	*ADCY3*
2q13	rs71801447	2	111925731	CTTATGTT /C	0.06	1.06(0.98-1.14)	1.6x10^{-01}	1.06(1.02-1.11)	2.5x10^{-03}	1.09(1.05-1.13)	7.7x10^{-06}	3.7x10^{-08}	*BCL2L11*
2q36.3	rs12479355	2	227226952	A/G	0.21	0.94(0.9-0.98)	2.5x10^{-03}	0.96(0.94-0.98)	8.8x10^{-04}	0.96(0.94-0.98)	4.7x10^{-04}	2.4x10^{-08}	-
3p13	rs6805189	3	71532113	T/C	0.48	0.96(0.92-0.99)	1.1x10^{-02}	0.97(0.95-0.99)	9.5x10^{-04}	0.97(0.95-0.99)	3.3x10^{-04}	4.6x10^{-08}	*FOXP1*
3p12.1	rs13066793	3	87037543	A/G	0.09	0.91(0.84-0.99)	2.8x10^{-02}	0.93(0.9-0.96)	1.7x10^{-05}	0.94(0.91-0.97)	1.5x10^{-04}	1.0x10^{-09}	*VGLL3*
3p12.1	rs9833888	3	99723580	G/T	0.22	1.06(1.01-1.1)	9.7x10^{-03}	1.03(1.01-1.06)	5.4x10^{-03}	1.06(1.04-1.08)	2.6x10^{-07}	5.2x10^{-10}	*CMSS1, FILIP1L*
3q23	rs34207738	3	141112859	CTT/C	0.41	1.04(1-1.07)	7.0x10^{-02}	1.05(1.03-1.07)	1.4x10^{-06}	1.06(1.04-1.08)	1.4x10^{-09}	3.2x10^{-15}	*ZBTB38*
3q26.31	rs58058861	3	172285237	G/A	0.21	1.05(1.01-1.1)	1.2x10^{-02}	1.03(1.01-1.05)	1.2x10^{-02}	1.06(1.04-1.09)	1.6x10^{-08}	1.9x10^{-10}	-

4p14	rs6815814	4	38816338	A/C	0.26	1.05(1-1.09)	2.8×10^{-02}	1.05(1.03-1.07)	2.2×10^{-05}	1.06(1.04-1.08)	6.1×10^{-08}	6.1×10^{-13}	-
4q21.23	4:84370124	4	84370124	TA/TAA	0.47	1.02(0.99-1.06)	2.1×10^{-01}	1.05(1.03-1.07)	3.6×10^{-06}	1.04(1.02-1.05)	1.7×10^{-04}	2.2×10^{-09}	*HELQ*
4q22.1	rs10022462	4	89243818	C/T	0.44	1.07(1.03-1.1)	3.5×10^{-04}	1.03(1.01-1.05)	6.3×10^{-03}	1.04(1.02-1.06)	9.4×10^{-06}	1.6×10^{-09}	*LOC105369192*
4q28.1	rs77528541	4	126843504	G/T	0.13	0.91(0.86-0.96)	6.3×10^{-04}	0.95(0.92-0.98)	1.2×10^{-03}	0.95(0.92-0.97)	4.8×10^{-05}	1.4×10^{-09}	-
5p15.33	rs116095464	5	345109	T/C	0.05	1.14(1.05-1.23)	1.5×10^{-03}	1.1(1.05-1.14)	1.8×10^{-05}	1.06(1.02-1.1)	2.6×10^{-03}	3.8×10^{-09}	*AHRR*
5q11.1	rs72749841	5	49641645	T/C	0.16	0.93(0.87-1)	3.7×10^{-02}	0.93(0.89-0.96)	1.9×10^{-04}	0.93(0.91-0.96)	8.5×10^{-06}	7.2×10^{-10}	-
5q11.1	rs35951924	5	50195093	A/AT	0.32	0.96(0.92-1)	4.4×10^{-02}	0.95(0.93-0.98)	5.6×10^{-05}	0.95(0.93-0.97)	4.0×10^{-07}	1.3×10^{-11}	-
5q22.1	rs6882649	5	111217786	T/G	0.34	0.94(0.91-0.98)	1.5×10^{-03}	0.96(0.94-0.98)	2.0×10^{-05}	0.97(0.95-0.99)	2.7×10^{-03}	3.7×10^{-09}	*NREP*
5q31.1	rs6596100	5	132407058	C/T	0.25	0.97(0.93-1.01)	1.2×10^{-01}	0.97(0.95-1)	2.9×10^{-02}	0.94(0.92-0.96)	5.2×10^{-08}	7.7×10^{-09}	*HSPA4*
5q35.1	rs4562056	5	169591487	G/T	0.33	1.04(1-1.08)	3.8×10^{-02}	1.03(1.01-1.06)	1.7×10^{-03}	1.05(1.03-1.07)	4.1×10^{-07}	4.7×10^{-10}	-
6p22.3	rs3819405	6	16399557	C/T	0.33	0.93(0.9-0.97)	6.9×10^{-04}	0.98(0.96-1)	8.5×10^{-02}	0.96(0.94-0.97)	2.2×10^{-06}	1.7×10^{-08}	*ATXN1*
6p22.3	rs2223621	6	20621238	C/T	0.38	1.05(1.02-1.09)	4.2×10^{-03}	1.04(1.02-1.06)	3.9×10^{-05}	1.04(1.02-1.06)	1.0×10^{-04}	3.0×10^{-10}	*CDKAL1*
6p22.2	rs71557345	6	26680698	G/A	0.07	0.92(0.86-0.98)	1.1×10^{-02}	0.92(0.89-0.96)	3.1×10^{-05}	0.92(0.88-0.96)	8.4×10^{-05}	3.9×10^{-10}	-
6q14.1	rs12207986	6	81094287	A/G	0.47	0.95(0.92-0.98)	3.9×10^{-03}	0.96(0.94-0.98)	9.6×10^{-05}	0.97(0.95-0.98)	2.0×10^{-04}	1.5×10^{-09}	-
6q23.1	rs6569648	6	130349119	T/C	0.24	0.91(0.88-0.95)	1.1×10^{-05}	0.97(0.95-0.99)	8.1×10^{-03}	0.94(0.92-0.96)	4.8×10^{-08}	3.0×10^{-12}	*L3MBTL3*
7p15.3	rs7971	7	21940960	A/G	0.35	0.97(0.94-1.01)	1.4×10^{-01}	0.97(0.95-0.99)	8.8×10^{-04}	0.96(0.94-0.98)	1.4×10^{-05}	1.9×10^{-08}	*DNAH11, CDCA7L*
7p15.1	rs17156577	7	28356889	T/C	0.11	1.11(1.04-1.18)	1.5×10^{-03}	1.06(1.03-1.09)	1.9×10^{-04}	1.05(1.02-1.08)	3.8×10^{-04}	4.3×10^{-09}	*CREB5*
7q21.3	rs17268829	7	94113799	T/C	0.28	1.07(1.03-1.11)	2.6×10^{-04}	1.05(1.02-1.07)	3.6×10^{-05}	1.05(1.03-1.07)	1.3×10^{-06}	4.5×10^{-13}	-
7q22.1	rs71559437	7	101552440	G/A	0.12	0.96(0.91-1.01)	1.0×10^{-01}	0.92(0.89-0.95)	2.5×10^{-06}	0.93(0.91-0.96)	9.1×10^{-07}	5.1×10^{-12}	*CUX1*
8q22.3	rs514192	8	102478959	T/A	0.32	1.06(1.02-1.1)	1.3×10^{-03}	1.03(1-1.05)	1.6×10^{-02}	1.05(1.03-1.07)	3.7×10^{-06}	5.6×10^{-09}	-
8q23.1	rs12546444	8	106358620	A/T	0.1	0.94(0.88-0.99)	3.1×10^{-02}	0.93(0.89-0.96)	3.1×10^{-05}	0.93(0.91-0.96)	5.8×10^{-06}	7.5×10^{-11}	*ZFPM3*
8q24.13	rs58847541	8	124610166	G/A	0.15	1.08(1.03-1.13)	1.7×10^{-03}	1.05(1.02-1.08)	7.8×10^{-04}	1.08(1.05-1.1)	7.3×10^{-09}	5.5×10^{-13}	-
9q33.1	rs1895062	9	119313486	A/G	0.41	0.97(0.94-1)	7.7×10^{-02}	0.97(0.95-0.99)	6.4×10^{-04}	0.94(0.92-0.95)	6.9×10^{-13}	1.1×10^{-14}	*ASTN2*
9q33.3	rs10760444	9	129396434	A/G	0.43	1.08(1.04-1.11)	3.2×10^{-05}	1.03(1.01-1.05)	4.9×10^{-03}	1.03(1.02-1.05)	2.8×10^{-04}	9.1×10^{-09}	*LMX1B*
9q34.2	rs8176636	9	136151579	I/D[10]	0.2	1.05(1-1.1)	5.4×10^{-02}	1.06(1.03-1.09)	2.5×10^{-06}	1.03(1.01-1.06)	3.2×10^{-03}	1.4×10^{-08}	*ABO*

10p14	rs67958007	10	9088113	TG/T	0.12	1.06(1-1.12)	$3.8x10^{-02}$	1.04(1.01-1.07)	$1.9x10^{-02}$	1.09(1.06-1.12)	$1.8x10^{-09}$	$1.7x10^{-10}$	-
10q23.33	rs140936696	10	95292187	C/CAA	0.18	1.07(1.02-1.12)	$6.1x10^{-03}$	1.05(1.02-1.08)	$5.0x10^{-04}$	1.04(1.02-1.07)	$7.4x10^{-04}$	$4.2x10^{-08}$	-
11p15	rs6597981	11	803017	G/A	0.48	0.96(0.93-1)	$3.3x10^{-02}$	0.96(0.94-0.97)	$5.0x10^{-06}$	0.96(0.94-0.97)	$5.7x10^{-07}$	$1.4x10^{-12}$	*PIDD1*
12q21.31	rs202049448	12	85009437	T/C	0.34	0.96(0.92-0.99)	$2.2x10^{-02}$	0.98(0.96-1)	$6.0x10^{-02}$	0.95(0.93-0.97)	$2.5x10^{-07}$	$2.7x10^{-08}$	-
12q24.31	rs206966	12	120832146	C/T	0.16	1.04(0.99-1.1)	$1.0x10^{-01}$	1.06(1.03-1.09)	$1.3x10^{-04}$	1.05(1.02-1.07)	$2.7x10^{-04}$	$3.8x10^{-08}$	-
14q32.33	rs10623258	14	105212261	C/CTT	0.45	1.06(1.01-1.1)	$9.8x10^{-03}$	1.03(1.01-1.05)	$3.7x10^{-03}$	1.04(1.02-1.06)	$2.7x10^{-05}$	$2.3x10^{-08}$	*ADSSL1*
16q12.2	rs28539243	16	54682064	G/A	0.49	1.05(1.01-1.09)	$1.2x10^{-02}$	1.05(1.03-1.07)	$1.3x10^{-06}$	1.05(1.03-1.07)	$3.6x10^{-08}$	$9.1x10^{-15}$	-
16q13	rs2432539	16	56420987	G/A	0.4	1.05(1.02-1.09)	$4.8x10^{-03}$	1.03(1.01-1.05)	$1.5x10^{-03}$	1.03(1.02-1.05)	$3.1x10^{-04}$	$4.0x10^{-08}$	*AMFR*
16q24.2	rs4496150	16	87085237	C/A	0.25	0.96(0.92-1)	$6.9x10^{-02}$	0.96(0.94-0.98)	$3.5x10^{-04}$	0.96(0.94-0.98)	$3.4x10^{-05}$	$8.1x10^{-09}$	-
17q21.2	rs72826962	17	40836389	C/T	0.01	0.99(0.81-1.2)	$8.9x10^{-01}$	1.23(1.12-1.35)	$2.6x10^{-05}$	1.2(1.11-1.3)	$5.1x10^{-06}$	$4.6x10^{-09}$	*CNTNAP1*
17q21.31	rs2532263	17	44252468	G/A	0.19	0.92(0.88-0.96)	$4.1x10^{-04}$	0.94(0.92-0.97)	$1.0x10^{-05}$	0.95(0.93-0.97)	$4.7x10^{-06}$	$6.9x10^{-13}$	*KANSL1*
18q12.1	rs117618124	18	29977689	T/C	0.05	0.86(0.79-0.94)	$6.5x10^{-04}$	0.93(0.88-0.97)	$2.8x10^{-03}$	0.89(0.85-0.92)	$4.5x10^{-08}$	$5.5x10^{-12}$	*GAREM1*
19p13.13	rs78269692	19	13158277	T/C	0.05	1.08(1-1.17)	$5.5x10^{-02}$	1.12(1.06-1.19)	$4.8x10^{-05}$	1.09(1.04-1.13)	$3.9x10^{-05}$	$1.9x10^{-09}$	*NFIX1*
19p13.12	rs2594714	19	13954571	G/A	0.23	0.94(0.9-0.98)	$1.7x10^{-03}$	0.95(0.93-0.97)	$1.6x10^{-05}$	0.97(0.95-0.99)	$6.7x10^{-03}$	$1.1x10^{-08}$	-
19p13.11	rs2965183	19	19545696	G/A	0.35	1.05(1.01-1.09)	$6.2x10^{-03}$	1.05(1.03-1.07)	$6.4x10^{-06}$	1.04(1.02-1.06)	$9.6x10^{-06}$	$6.3x10^{-12}$	*GATAD2A, MIR640*
19q13.22	rs71338792	19	46183031	A/AT	0.23	1.04(1-1.09)	$6.5x10^{-02}$	1.05(1.02-1.08)	$6.6x10^{-04}$	1.05(1.03-1.07)	$8.1x10^{-06}$	$3.5x10^{-09}$	*GIPR*
20p12.3	rs16991615	20	5948227	G/A	0.06	1.09(1.02-1.17)	$1.8x10^{-02}$	1.05(1.01-1.09)	$1.5x10^{-02}$	1.1(1.06-1.14)	$1.4x10^{-07}$	$1.9x10^{-09}$	*MCM8*
20q13.13	rs6122906	20	48945911	A/G	0.18	1.08(1.03-1.13)	$6.3x10^{-04}$	1.05(1.02-1.07)	$3.8x10^{-04}$	1.05(1.03-1.07)	$2.9x10^{-05}$	$2.5x10^{-10}$	-
22q13.1	rs738321	22	38568833	C/G	0.38	0.94(0.91-0.97)	$5.1x10^{-04}$	0.96(0.94-0.98)	$1.7x10^{-04}$	0.95(0.93-0.97)	$2.7x10^{-08}$	$1.0x10^{-13}$	*PLA2G6*
22q13.2	rs73161324	22	42038786	C/T	0.06	1.14(1.05-1.25)	$2.7x10^{-03}$	1.11(1.06-1.16)	$1.4x10^{-06}$	1.06(1.02-1.09)	$3.8x10^{-03}$	$2.0x10^{-09}$	*XRCC6*
22q13.31	rs28512361	22	46283297	G/A	0.11	1.06(0.99-1.14)	$8.4x10^{-02}$	1.08(1.04-1.13)	$2.0x10^{-05}$	1.05(1.02-1.08)	$5.7x10^{-04}$	$2.3x10^{-08}$	-

[1] The most significant variant at each locus is shown.

[2] Chromosome

631 [3] Build 37 position

632 [4] Major/minor allele (forward strand)

633 [5] Minor allele frequency in controls in OncoArray dataset

634 [6] Per-allele odds ratio (95% confidence limits)

635 [7] *P*-value (see Online Methods)

636 [8] Genes within 2kb

637 [9] 21 base-pair deletion

638 [10] 36 base-pair deletion

639

640

641

642

643

644 **Figure Legends**

645 **Figure 1. (a)** Manhattan plot showing $\log_{10}P$-values for SNP associations with overall breast cancer
646 **(b)** Manhattan plot after excluding previously identified associated regions. The red line denotes
647 "genome-wide" significance ($P<5\times10^{-8}$); the blue line denotes $P<10^{-5}$.

648

649 **References**

650 1 Amos, C. I. *et al.* The OncoArray Consortium: a Network for Understanding the Genetic
651 Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev*, doi:10.1158/1055-
652 9965.EPI-16-0106 (2016).
653 2 Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast
654 cancer risk. *Nature genetics* **45**, 353-361, doi:10.1038/ng.2563 (2013).
655 3 Long, J. *et al.* Genome-wide association study in east Asians identifies novel susceptibility
656 loci for breast cancer. *PLoS Genet* **8**, e1002532, doi:10.1371/journal.pgen.1002532 (2012).
657 4 Cai, Q. *et al.* Genome-wide association analysis in East Asians identifies breast cancer
658 susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nature genetics* **46**, 886-890,
659 doi:10.1038/ng.3041 (2014).
660 5 Long, J. *et al.* A common deletion in the APOBEC3 genes and breast cancer risk. *J Natl Cancer*
661 *Inst* **105**, 573-579, doi:10.1093/jnci/djt018 (2013).
662 6 He, C. *et al.* Genome-wide association studies identify loci associated with age at menarche
663 and age at natural menopause. *Nature genetics* **41**, 724-728, doi:10.1038/ng.385 (2009).
664 7 Kawase, T. *et al.* PH domain-only protein PHLDA3 is a p53-regulated repressor of Akt. *Cell*
665 **136**, 535-550, doi:10.1016/j.cell.2008.12.002 (2009).
666 8 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome
667 sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).
668 9 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast
669 tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).
670 10 Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*
671 **163**, 506-519, doi:10.1016/j.cell.2015.09.033 (2015).
672 11 Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic
673 and transcriptomic landscapes. *Nat Commun* **7**, 11479, doi:10.1038/ncomms11479 (2016).
674 12 Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted
675 gene functions. *Nat Commun* **6**, 5890, doi:10.1038/ncomms6890 (2015).
676 13 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-
677 based method for gene-set enrichment visualization and interpretation. *PloS one* **5**, e13984,
678 doi:10.1371/journal.pone.0013984 (2010).
679 14 Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. *Nat*
680 *Rev Cancer* **10**, 116-129, doi:10.1038/nrc2780 (2010).
681 15 Heldin, C. H. Targeting the PDGF signaling pathway in tumor treatment. *Cell Commun Signal*
682 **11**, 97, doi:10.1186/1478-811X-11-97 (2013).
683 16 Howe, L. R. & Brown, A. M. Wnt signaling and breast cancer. *Cancer Biol Ther* **3**, 36-41
684 (2004).
685 17 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
686 association summary statistics. *Nature genetics* **47**, 1228-1235, doi:10.1038/ng.3404 (2015).
687 18 Lin, Y., Ma, W. & Benchimol, S. Pidd, a new death-domain-containing protein, is induced by
688 p53 and promotes apoptosis. *Nature genetics* **26**, 122-127, doi:10.1038/79102 (2000).
689 19 Fox, S. B. *et al.* CITED4 inhibits hypoxia-activated transcription in cancer cells, and its
690 cytoplasmic location in breast cancer is associated with elevated expression of tumor cell
691 hypoxia-inducible factor 1alpha. *Cancer research* **64**, 6075-6081, doi:10.1158/0008-
692 5472.CAN-04-0708 (2004).

693

694

695 **Online Methods**

696 Details of the studies and genotype calling and quality control (QC) for the iCOGS and eleven other
697 GWAS are described elsewhere[2,20]. Seventy-eight studies participated in the breast cancer
698 component of the OncoArray, of which 67 studies contributed European ancestry data and 12
699 contributed Asian ancestry data (one study, NBCS, was excluded as there were no controls from
700 Norway) (**Supplementary Table 1**). The majority of studies were population-based case-control
701 studies, or case-control studies nested within population-based cohorts, but a subset of studies
702 oversampled cases with a family history of the disease. All studies provided core data on disease
703 status and age at diagnosis/observation, and the majority provided additional data on clinico-
704 pathological factors and lifestyle factors, which have been curated and incorporated into the BCAC
705 database (version 6). All participating studies were approved by their appropriate ethics review
706 board and all subjects provided informed consent.

707 *OncoArray SNP Selection*

708 Approximately 50% of the SNPs for the OncoArray were selected as a “GWAS backbone” (Illumina
709 HumanCore), which aimed to provide high coverage for the majority of common variants through
710 imputation. The remaining SNPs were selected from lists supplied by each of six disease-based
711 consortia, together with a seventh list of SNPs of interest to multiple disease-focused groups.
712 Approximately 72k SNPs were selected specifically for their relevance to breast cancer. These
713 included: (a) SNPs showing evidence of association from previous genotype data, based on a
714 combined analysis of eleven existing GWAS together the data from the iCOGS experiment; (b) SNPs
715 showing evidence of association with ER-negative disease (through a combined analysis with the
716 CIMBA consortium), triple negative disease, breast cancer diagnosed before age 40 years, high grade
717 disease, node positive disease or ductal carcinoma-in-situ; (c) SNPs potentially associated with
718 breast cancer survival; (d) SNPs selected for fine-mapping of 55 regions showing evidence of breast
719 cancer association at genome-wide significance; (e) rare variants showing evidence of association
720 through exome sequencing in multiple case families, whole-genome sequencing in high-risk cases
721 (DRIVE), or analysis of the ExomeChip (BCAC); (f) specific follow-up of regions of interest from breast
722 cancer GWAS in Asian, Latina and African/African-American women; (g) SNPs associated with breast
723 density, selected from GWAS conducted by the MODE consortium; (h) breast tissue-specific eQTLs (i)
724 lists of functional candidates from >30 groups. Lists were merged with lists from the other consortia
725 as described elsewhere[1].

726 *OncoArray Calling and QC*

Of the 568,712 variants selected for genotyping, 533,631 were successfully manufactured on the array (including 778 duplicate probes). Genotyping for the breast cancer component of the OncoArray, which included 152,492 samples, was conducted at six sites. Details of the genotyping calling for the OncoArray are described in more detail elsewhere[1]. Briefly, we developed a single calling pipeline that was applied to more than 500,000 samples. An initial cluster file was generated using data from 56,284 samples, selected to cover all the major genotyping centres and ethnicities, using the Gentrain2 algorithm. Variants likely to have problematic clusters were selected for manual inspection using the following criteria: call rate below 99%, variants with minor allele frequency (MAF)<0.001, poor Illumina intensity and clustering metrics, or deviation from the expected frequency as observed in the 1000 Genomes Project. This resulted in manual adjustment of the cluster file for 3,964 variants, and the exclusion of 16,526 variants. The final cluster file was then applied to the full dataset.

We excluded probable duplicates and close relatives within each study, and probable duplicates across studies. We excluded samples with a call rate <95% or samples with extreme heterozygosity (4.89 SD from the mean for the ethnicity). Ancestry was computed using a principal component analysis, applied to the full OncoArray dataset, using 2318 informative markers on a subset of ~47,000 samples. The analysis presented here was restricted to women of European ancestry, defined as individuals with an estimated proportion of European ancestry >0.8, and women of East Asian ancestry (estimated proportion of Asian ancestry >0.4), with reference to the HapMap (v2) populations, based on the first two principal components. After quality control exclusions and removing overlaps with the previous iCOGS and GWAS genotyping used in the analysis, the final dataset comprised data from 61,282 cases and 45,494 of European ancestry 7,799 cases and 6,480 controls of Asian ancestry.

We excluded SNPs with a call rate <95% in any consortium, SNPs not in Hardy-Weinberg equilibrium ($P<10^{-7}$ in controls or $P<10^{-12}$ in cases) and SNPs with concordance <98% among 5,280 duplicate sample pairs. For the imputation, we additionally excluded SNPs with a MAF<1% and a call rate <98% in any consortium, SNPs that could not be linked to the 1000 Genomes Project reference or differed significantly in frequency from the 1000 Genomes Project dataset (using the criterion $\frac{(p_1-p_0)^2}{((p_1+p_0)(2-p_1-p_0))} > 0.007$, where p_0 and p_1 are the MAFs in the 1000 Genomes Project and OncoArray European datasets, respectively). A further 1,128 SNPs where the cluster plot was judged to be not ideal on visual inspection were excluded. Of the 533,631 SNPs that were manufactured on the array, 494,763 SNPs passed the initial QC and 469,364 SNPs were used in the imputation.

Genotype Imputation

760 All samples were imputed using the October 2014 (version 3) release of the 1000 Genomes Project
761 dataset as the reference panel and number of sampled haplotypes per individual (Nhap)=800. The
762 iCOGS, OncoArray and nine of the GWAS datasets were imputed using a two-stage imputation
763 approach, using SHAPEIT2 for phasing and IMPUTEv2 for imputation[21,22]. The imputation was
764 performed in 5Mb non-overlapping intervals. The subjects were split into subsets of ~10,000
765 samples; where possible subjects from the same study were included in the same subset. The BPC3
766 and EBCG studies were imputed separately using MACH and Minimac[23,24]. 99.6% of SNPs with
767 frequency >1% were imputable with r^2>0.3 in the OncoArray dataset and 99.1% in the iCOGS
768 dataset. We generated estimated genotypes for all SNPs that were polymorphic (MAF>0.1%) in
769 either European or Asian samples (~21M SNPs). For the current analysis, however, we restricted to
770 SNPs with MAF>0.5% in the European OncoArray dataset (11.8M SNPs). One-step imputation
771 (without pre-phasing) was performed, on the iCOGS and OncoArray datasets, as a quality control
772 step for those associated loci where the imputation quality score was <0.9. Imputation quality for
773 the lead variants, as assessed by the IMPUTE2 quality score in the OncoArray dataset, was >0.80 for
774 all but one locus (**Supplementary Table 27**) rs72749841, quality score=0.65).

775

776 *Principal Components Analysis*

777 To adjust for potential (intra-continental) population stratification in the OncoArray dataset,
778 principal components analysis was performed using data from 33,661 uncorrelated SNPs (which
779 included 2,318 SNPs specifically selected on informativeness for determining continental ancestry)
780 with a MAF of at least 0.05 and maximum correlation of 0.1 in the OncoArray dataset, using
781 purpose-written software (http://ccge.medschl.cam.ac.uk/software/pccalc). For the main analyses,
782 we used the first ten principal components, as additional components did not further reduce
783 inflation in the test statistics. We used nine principal components for the iCOGS and up to ten
784 principal components for the other GWAS, where this was found to reduce inflation.

785 *Statistical Analyses*

786 Per-allele ORs and standard errors were generated for the OncoArray, iCOGS and each GWAS,
787 adjusting for principal components using logistic regression. The OncoArray and iCOGS analyses
788 were additionally adjusted for country and study, respectively. For the OncoArray analysis, we
789 adjusted for country and 10 principal components. Adjustment for country rather than study was
790 used to improve power since some studies had no few or no controls. We evaluated the adequacy of
791 this approach by comparing the inflation in the test statistic with that obtained in corresponding

792 analysis in which we adjusted for study – the inflation was very similar (λ=1.15 vs. 1.17, based on the
793 backbone SNPs, equivalent to λ_{1000}=1.003, for a study of 1,000 cases and 1,000 controls, in both
794 cases). As an additional sensitivity analysis, we computed the effect sizes for the 65 novel loci
795 adjusting for study – the effect sizes were essentially identical to those presented. Estimates were
796 derived using ProbAbel for the BPC3 and EBCG studies[25], SNPTEST for the remaining GWAS and
797 purpose written software for the iCOGS and OncoArray datasets. OR estimates and standard errors
798 were combined in a fixed effects inverse variance meta-analysis using METAL[26], adjusting the GWAS
799 (but not iCOGS or OncoArray) results for genomic control as described previously[2]. For the GWAS,
800 results were included in the analysis for all SNPs with MAF>0.01 and imputation r^2>0.3. For iCOGS
801 and OncoArray we included all SNPs with r^2>=0.3 and MAF>0.005 (11.8M SNPs in total). We viewed
802 the primary tests of association as those based on all the meta-analysis over all stages, as this has
803 been shown to be powerful than tests based on a test-replication approach[27]. Eight sets of variants
804 were associated with breast cancer at $P<5\times10^{-8}$ but were close to previous susceptibility regions, and
805 these became non-significant after adjustment for the previously identified lead variant. Two SNPs
806 on 22q13.2, rs141447235 and rs73161324, were both associated with overall breast cancer but,
807 despite lying >500kb apart, were strongly correlated with each other (r^2=0.50) and hence were
808 considered as a single novel signal.

809

810 For SNPs showing evidence of association, we additionally computed genotype-specific ORs for the
811 iCOGS and OncoArray dataset, and per-allele ORs for ER-negative and ER-positive disease.
812 Departures from a log-additive model were evaluated using a one degree of freedom likelihood ratio
813 test, comparing the log-additive model (genotypes parametrised as the number of rare alleles
814 carried) with the general model estimating ORs for each genotype. The genotype-specific risks for all
815 variants were consistent with a log-additive model **(**P>0.01; **Supplementary Table 28)**. Tests for
816 differences in the OR by ER-status were derived using case-only analyses, in which estimates were
817 derived by logistic regression separately in the iCOGS and OncoArray datasets, adjusted as before,
818 and then combined in a fixed-effects meta-analysis. These analyses were performed in R[28].

819 We assessed heterogeneity in the OR estimates among studies within each of the OncoArray, iCOGS
820 and GWAS components, and between the (combined) estimates for the three components, using
821 both the I^2 statistic and the P-value for Cochran's Q statistic (**Supplementary Table 27**). There was no
822 evidence of heterogeneity among studies in the ORs for any of the loci in the OncoArray, but three
823 loci showed some evidence of heterogeneity in the ORs among the GWAS, iCOGS and OncoArray
824 datasets.

825 To determine whether there were multiple independent signals in a given region, we performed
826 multiple logistic regression analysis using SNPs within 500kb of each lead SNP, adjusting for the lead
827 SNP. We used the genotypes derived by one-step imputation, performed the analyses separately in
828 the iCOGS and Oncoarray datasets and combined the results (adjusted effect sizes and standard
829 errors) using a fixed effects meta-analysis. For one of the two loci for which there was an additional
830 signal significant at $P<5\times10^{-8}$, the lead SNP from the one-step imputation differed from the lead SNP
831 in the overall analysis, but was strongly correlated with it (**Supplementary Table 8**).

832

833 *Definition of Known Hits*

834 We attempted to identify all associations previously reported from genome-wide or candidate
835 analysis at a significance level $P<5\times10^{-8}$ for overall breast cancer, ER-negative or ER-positive breast
836 cancer, in *BRCA1* or *BRCA2* carriers, or in meta-analyses of these categories. Where multiple studies
837 reported associations in the same region, we used the first reported association unless later studies
838 identified a variant that was clearly more strongly associated. We only included one SNP per 500kb
839 interval, unless joint analysis provided clear evidence ($P<5\times10^{-8}$) of more than one independent
840 signal. For the analysis of credible risk variants (CRVs), we restricted attention to regions where the
841 most significant signal had a P-value$<10^{-7}$ in Europeans (77 regions). To avoid complications with
842 defining CRVs for secondary signals, we considered only the primary signal and defined CRVs as
843 those whose *P*-value was within two orders of magnitude of the most significant *P*-value.

844 *In-Silico Analysis of CRVs*

845 We combined multiple sources of *in silico* functional annotation from public databases to help
846 identify potential functional SNPs and target genes. To investigate functional elements enriched
847 across the region encompassing the strongest CRVs, we analysed chromatin biofeatures data from
848 the Encyclopedia of DNA Elements (ENCODE) Project[29], Roadmap Epigenomics Projects[30] and other
849 data obtained through the National Center for Biotechnology Information (NCBI) Gene Expression
850 Omnibus (GEO) namely: Chromatin State Segmentation by Hidden Markov Models (chromHMM),
851 DNase I hypersensitive and histone modifications of epigenetic markers H3K4, H3K9, and H3K27 in
852 Human Mammary Epithelial (HMEC) and myoepithelial (MYO) cells, T47D and MCF7 breast cancer
853 cells and TF ChIP-seq in a range of breast cell lines (**Supplementary Table 12**).

854 *Association of Genomic Features with CRVs*

855 We first defined credible candidate variants as those located within 500kb of the most significant
856 SNP in each region, and with *P*-values within two orders of magnitude of the most significant SNPs.
857 This is approximately equivalent to flagging variants whose posterior probability of causality is within
858 two orders of magnitude of that of the most significant SNP[31,32]. We then selected 800 random 1Mb
859 control regions separated by at least 1Mb from each other and from the intervals defined by the
860 associated SNPs. The association with each feature was then evaluated using logistic regression, with
861 being a CRV as the outcome, and adjusting for the dependence due to linkage disequilibrium using
862 robust variance estimation, clustering on region, using the R package multiwayvcov.

863 *eQTL analyses*

864 Expression QTL analyses were performed using data from The Cancer Genome Atlas (TCGA) and
865 Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) projects[9,33]. The TCGA
866 eQTL analysis was based on 458 breast tumours that had matched gene expression, copy number,
867 and methylation profiles together with the corresponding germline genotypes available. All 458
868 individuals were of European ancestry as ascertained using the genotype data and the Local
869 Ancestry in adMixed Populations (LAMP) software package (LAMP estimate cut-off >95%
870 European)[34]. Germline genotypes were imputed into the 1000 Genomes Project reference panel
871 (October 2014 release) using IMPUTE2[23,35]. Gene expression had been measured on the Illumina
872 HiSeq 2000 RNA-Seq platform (gene-level RSEM normalized counts[36]), copy number estimates were
873 derived from the Affymetrix SNP 6.0 (somatic copy number alteration minus germline copy number
874 variation called using the GISTIC2 algorithm[37]), and methylation beta values measured on the
875 Illumina Infinium HumanMethylation450. Expression QTL analysis focused on all variants within 500
876 kb of the most significantly associated risk SNP in 142 genomic regions (each 2-Mb wide) containing
877 at least one previously identified or new overall breast cancer risk locus confirmed at genome-wide
878 significance in the current meta-analysis. Each variant was evaluated for its association with the
879 expression of every gene within 2 Mb that had been profiled for each of the three data types. The
880 effects of tumour copy number and methylation on gene expression were first regressed out using a
881 method described previously[38]. eQTL analysis was performed by linear regression, with residual gene
882 expression as outcome, germline SNP genotype dosage as the covariate of interest and *ESR1*
883 expression and age as additional covariates, using the R package Matrix eQTL[39].

884 The METABRIC eQTL analysis was based on 138 normal breast tissue samples resected from breast
885 cancer patients of European ancestry. Germline genotyping for the METABRIC study was also done
886 on the Affymetrix SNP 6.0 array, and gene expression in the METABRIC study was measured using
887 the Illumina HT12 microarray platform (probe-level estimates). No adjustment was implemented for

888 somatic copy number and methylation status since we were evaluating eQTLs in normal breast
889 tissue. All other steps were identical to the TCGA eQTL analysis described above.
890

891 *INQUISIT*

892 We developed a computational pipeline, **integrated expression quantitative trait and in silico
893 prediction of GWAS targets (INQUISIT), to interrogate publically available data for the prioritisation
894 of candidate target genes.
895

896 *Data used for INQUISIT*: Chromatin interaction data from ENCODE ChIA-PET analysis in MCF-7 cells
897 for RNApolII, ERalpha, and CTCF factors were downloaded using UCSC Table Browser[40]. Hi-C data
898 derived from HMECs were obtained from Rao *et al*.[41], using "interaction loops" as defined in the
899 publication. Data were reformatted to facilitate intersection of query SNPs using BEDTools
900 "intersect"[42]. For all interactions, termini were intersected with promoters using GENCODE v19[43]
901 Basic gene annotations, where we defined promoters as -1.0 kb - +0.1 kb surrounding a transcription
902 start site.
903

904 Enhancer-target gene predictions by several computational algorithms were collected. Each of these
905 datasets assigns genes to enhancers. We used all MCF-7 and HMEC enhancer predictions (low and
906 high stringency) made by PreSTIGE[44], IM-PET enhancer-gene predictions in MCF-7, HMEC and
907 HCC1954 cell lines[45]. Enhancer-transcription start site (E-TSS) links were identified from the
908 FANTOM5 Consortium were identified[46], and enhancers detected in mammary epithelial cells were
909 intersected with E-TSS links. We also collected typical and super-enhancers in MCF-7, HMEC and
910 HCC1954 cells defined by Hnisz *et al*.[47].
911

912 TF ChIP-seq peak data for ESR1, FOXA1, GATA3, TCF7L2 and E2F1 from MCF-7, T47D and MCF-10A
913 cells were downloaded in narrowPeak format from ENCODE. H3K4me3 and H3K9ac (characteristic of
914 promoters) histone modification ChIP-seq peak data for all breast cells were obtained from ENCODE
915 and Roadmap Epigenomics Project. ChromHMM data for breast cell samples (HMEC and
916 myoepithelial: E027, E028 and E119) were downloaded from Roadmap Epigenomics.
917

918 Expression QTL analyses were conducted as described above. In the interpretation of the eQTL
919 results for INQUISIT (and in general) we focused on the overlap between the CRVs (risk signal) and
920 the top eQTL variants for a given gene (eQTL signal). If the eQTL *P*-value for a CRV was the same as,
921 or within 1/100th of the eQTL *P*-value of the SNP most significantly associated with expression of a

922 particular gene, that gene and the corresponding CRV were assigned a point for being an eQTL in
923 INQUISIT.
924
925 Topologically-associated domain (TAD) boundaries were derived from Hi-C data[41]. Genomic intervals
926 corresponding to "contact domains" from eight human cell types were merged using BEDTools
927 "merge" resulting in annotation of regions most likely to encompass TAD units. Inter-TAD
928 boundaries were identified using BEDTools "complement".
929
930 Gene level RNA-seq expression data generated under multiple experimental conditions in MCF-7 and
931 normal mammary epithelial cells were downloaded from ENCODE. The FPKM (Fragments Per
932 Kilobase of exon per Million fragments Mapped) values for each gene were extracted using the
933 metagene R package[48] and averaged across all experiments to give an approximation of expression
934 in breast cells. Accession numbers are given in **Supplementary Table 29**.
935
936 *INQUISIT pipeline*
937 Candidate target genes were evaluated by assessing each CRV's potential impact on regulatory or
938 coding features. Scores categorised by 1) distal gene regulation, 2) proximal gene regulation, or 3)
939 impact on protein coding were calculated using the following criteria (see also **Supplementary Table**
940 **16**).
941
942 Genomic annotation data for target gene predictions (chromatin interaction and computational
943 enhancer-promoter assignment), ChIP-seq, histone modification, and chromHMM were curated into
944 a BED formatted database. We intersected the chromosomal positions of CRVs with each category of
945 genomic annotation data using BEDTools "intersect" (minimum 1 bp overlap), resulting in
946 annotation of SNP-gene pairs with presence or absence of multiple classes of genomic data. Each
947 gene was scored using a custom R script on the basis of the following criteria:
948 - For distally regulated genes, a candidate gene was given 2 points if a CRV fell in an element that
949 revealed long range ChIA-PET or Hi-C interactions with that gene's promoter. One point was
950 added to a gene's score in the case of enhancers predicted by computational methods to target
951 that gene (in addition to experimental interactions if also observed). If the distal elements
952 harbouring SNPs also overlapped enriched cistromic TF (ESR1, FOXA1, GATA3, TCF7L2, E2F1)
953 ChIP-seq peaks, an additional point was given when one SNP-Enhancer-ChIP-seq peak
954 intersection occurred, but two points when there were multiple TF binding sites overlapping SNPs
955 in distinct interactions or enhancers (see **Supplementary Table 16** for details). One point was

956 given to significant eSNP-eGENE pairs. Predicted distal target genes which were among the list of
957 breast cancer driver genes were up-weighted with a further point (except for the analysis of
958 driver gene enrichment). Information regarding TAD boundaries was used to down-weight genes:
959 genes which were separated from CRVs by a TAD boundary were down-weighted by multiplying
960 their scores by 0.05. Scores for genes exhibiting no expression in MCF7 or HMEC (mean FPKM =
961 0) were multiplied by 0.1. This resulted in scores for each candidate target gene ranging from 0 to
962 8.
963 - Variants were treated as potentially affecting proximal promoter regulation if they resided
964 between -1.0 and +0.1 kb surrounding a transcription start site. Additional points was awarded to
965 genes when variants overlapped promoter H3K4me3 or H3K9ac histone modification peaks,
966 intersected with ESR1, FOXA1, GATA3, TCF7L2 or E2F1 TF binding sites, were significant eSNP-
967 eGENE pairs, and if the gene was annotated as a breast cancer driver gene. Gene scores were
968 down-weighted (by a factor of 0.1) if they lacked expression in MCF-7 or HMEC samples.
969 Resultant scores ranged from 0 to 5.
970 - Intragenic variants were evaluated for their potential to impact protein function using a range of
971 in silico prediction tools (CADD[49], FATHMM[50], LRT[51], MutationAssessor[52], Mutation Taster 2[53],
972 PolyPhen-2[54], PROVEAN[55] and SIFT[56] for missense variants; Human Splicing Finder[57] and
973 MaxEntScan[58] for splice variants). We scored genes with missense and nonsense variants
974 predicted to be functionally deleterious, and points for genes harbouring variants predicted to
975 alter splicing. Genes could therefore carry SNPs which affect coding and splicing and receive
976 increased scores. Additional points were given to genes which were breast cancer driver genes.
977 We multiplied scores by 0.1 when genes showed a lack of expression in breast cells. Possible
978 coding scores ranged from 0-4.
979
980 *Enrichment of Somatic Breast Cancer Driver Genes in INQUISIT Target Gene Predictions*
981 We listed 147 unique protein coding driver genes for breast cancer identified from four recent
982 tumour genome and exome sequencing studies (considering *ZNF703* and *FGFR1* as independent
983 genes; **Supplementary Table 30**)[8-11]. First, we examined overlap between this list of 147 genes and
984 the total set of unique target genes predicted by INQUISIT (n = 689) by one or more of the three
985 regulatory mechanisms (distal, promoter, and coding). The significance of this overlap was assessed
986 by randomly drawing (without replacement) 689 genes from the set of all protein coding genes
987 (GENCODE release 19, n = 20,243) one million times and calculating the probability of observing the
988 same (or stronger) overlap with the list of 147 drivers. Second, we hypothesised that this enrichment
989 would be stronger with progressively higher INQUISIT scores. We categorised all 20,243 protein

coding genes into four levels based on their INQUIST scores (level 1: coding score 2, promoter score 3-4, distal score >4; level 2: coding 1, promoter 1-2, distal 1-4; level 3: any score >0 but <1; level 4: score 0 i.e. not a predicted target). The gene nearest to a risk locus is frequently assigned as a candidate target gene in GWAS in the absence of additional functional analysis[59]. We observed that seven of the 147 drivers were among the genes nearest to a previously or newly identified breast cancer risk locus. Therefore, we used logistic regression, including data for all target genes predicted by INQUISIT, with driver status as outcome, and evaluated INQUISIT score level and nearest gene status as potential predictors of driver status (**Supplementary Table 20**).

Lead SNPs at 142 breast cancer risk associated loci were used as input into DEPICT which was then run using the default settings[12]. We examined the relative performance of INQUISIT and DEPICT in predicting driver gene status using logistic regression models as above (**Supplementary Table 20**), adding DEPICT prediction as a covariate.

Chromatin Conformation Capture (3C)

MCF7 (ATCC #HTB22) and MDA-MB-231 (ATCC #HTB26) breast cancer cell lines were grown in RPMI medium with 10% FCS and antibiotics. Bre-80 normal breast epithelial cells (provided as a gift from Roger Reddel, CMRI, Sydney) were grown in DMEM/F12 medium with 5% horse serum (HS), 10 µg/ml insulin, 0.5 µg/ml hydrocortisone, 20 ng/ml epidermal growth factor, 100 ng/ml cholera toxin and antibiotics. Cell lines were maintained under standard conditions, routinely tested for *Mycoplasma* and short tandem repeat (STR) profiled to confirm cell line identity. 3C libraries were generated using *Eco*RI as described previously[60]. 3C interactions were quantitated by real-time PCR (qPCR) using primers designed within restriction fragments (**Supplementary Table 31**). qPCR was performed on a RotorGene 6000 using MyTaq HS DNA polymerase (Bioline) with the addition of 5 mM of Syto9, annealing temperature of 66°C and extension of 30 sec. 3C analyses were performed in three independent 3C libraries from each cell line with each experiment quantified in duplicate. BAC clones covering each region were used to create artificial libraries of ligation products in order to normalize for PCR efficiency. Data were normalized to the signal from the BAC clone library and, between cell lines, by reference to a region within *GAPDH*. All qPCR products were electrophoresed on 2% agarose gels, gel purified and sequenced to verify the 3C product.

Plasmid Construction and Reporter Assays

Promoter-driven luciferase reporter constructs were generated by insertion of PCR amplified fragments or synthesised gBlocks (Integrated DNA Technologies) containing the *KLHDC7A*, *PIDD1* or

1023 *CITED4* promoters into the *Kpn*I/*Hind*III sites of pGL3-Basic. For the 1p34 locus, a 1169 bp putative
1024 regulatory element (PRE1) or 951 bp PRE2 were synthesised as gBlocks and cloned into the
1025 *BamH*I/*Sal*I sites of the *CITED4*-promoter construct. The minor alleles of SNPs were introduced into
1026 promoter or PRE sequences by overlap extension PCR or gBlocks. Sequencing of all constructs
1027 confirmed variant incorporation (AGRF). MCF7 or Bre-80 cells were transfected with equimolar
1028 amounts of luciferase reporter plasmids and 50 ng of pRLTK transfection control plasmid with
1029 Lipofectamine 2000. The total amount of transfected DNA was kept constant at 600 ng for each
1030 construct by the addition of pUC19 as a carrier plasmid. Luciferase activity was measured 24 hr
1031 posttransfection by the Dual-Glo Luciferase Assay System. To correct for any differences in
1032 transfection efficiency or cell lysate preparation, *Firefly* luciferase activity was normalized to *Renilla*
1033 luciferase, and the activity of each construct was measured relative to the reference promoter
1034 constructs, which had a defined activity of 1. Statistical significance was tested by log transforming
1035 the data and performing 2-way ANOVA, followed by Dunnett's multiple comparisons test in
1036 GraphPad Prism.
1037
1038 *Global Genomic Enrichment Analyses*
1039 We performed stratified LD score regression analyses[17] for overall breast cancer as well as stratified
1040 by ER status using the summary statistics based on the meta-analyses of the OncoArray, GWAS and
1041 iCOGS datasets. We restricted analysis to all SNPs present on the HapMap version 3 dataset that had
1042 a MAF > 1% and an imputation quality score R^2>0.3 in the OncoArray data. LD scores were calculated
1043 using the 1000 Genomes Project Phase 3 EUR reference panel.

1044 We first created a "full baseline model" as previously described that included 24 non-cell type
1045 specific publicly available annotations as well as 24 additional annotations that included a 500-bp
1046 window around each of the 24 main annotations[17]. Additionally, we also included 100-bp windows
1047 around ChIP-seq peaks as well as one annotation containing all SNPs leading to a total of 53
1048 overlapping annotations.

1049 We subsequently performed analyses using cell-type specific annotations for four histone marks
1050 H3K4me1, H3K4me3, H3K9ac and H3K27ac across 27-81 cell types depending on histone mark[17].
1051 Each cell-type-specific annotation corresponded to a histone mark in a single cell type, and there
1052 were 220 such annotations in total. We augmented the baseline model by adding these annotations
1053 individually, creating 220 separate models, each with 54 annotations (53+1). This procedure controls
1054 for the overlap with the 53 functional categories in the full baseline model but not with the 219
1055 other cell type specific annotations.

1056 We further tested the differences in functional enrichment between ER-positive and ER-negative
1057 subsets through a Wald test, using the regression coefficients and standard errors for the two
1058 subsets based on the models described above.

1059

1060 *Contribution of Identified Variants to the Familial Relative Risk of Breast Cancer*

1061 We estimated the proportion of the familial risk of breast cancer due to the identified variants,
1062 under a log-additive model, using the formula:

1063 $\sum_i p_i(1-p_i)(\beta_i^2-\tau_i^2)/\ln(\lambda))$, where p_i is the MAF for variant *i*, $\beta_i$is the log(OR) estimate for
1064 variant *i*, τ_i is the standard error of β_i and $\lambda=2$ is the assumed overall familial relative risk.

1065

1066 To compute the corresponding estimate for the FRR due to all variants, we wish to estimate
1067 $h_f^2=\sum_i 2p_i(1-p_i)\beta_i^2$ where the sum is now over the all variants and β_i is the true relative risk
1068 conferred by variant *i*, assuming a log-additive model. We refer to h_f^2 as the *frailty scale* heritability.
1069 We first obtained the estimated observed heritability based on the full set of summary estimates
1070 using LD Score Regression[17] and then converted this to an estimate on the frailty scale using the
1071 $h_f^2 = {h_{obs}^2}/{P(1-P)}$, where *P* is the proportion of samples in the population that are cases.

1072

1073 *Pathway Analyses*

1074 The pathway gene set database (http://download.baderlab.org/EM_Genesets, file
1075 Human_GOBP_AllPathways_no_GO_iea_April_01_2017_symbol.gmt)[13] from the Bader lab dated
1076 April 1, 2017 was used in all analyses. This database contains pathways from Reactome[61], NCI
1077 Pathway Interaction Database[62], GO (Gene Ontology) biological process[63], HumanCyc[64], MSigdb[65],
1078 NetPath[66] and Panther[67]. For GO, terms inferred from electronic annotation were excluded from our
1079 analyses. The same pathway may be defined in two or more databases with potentially different sets
1080 of genes. All versions of such 'duplicate' pathways were included. To provide more biologically
1081 meaningful results and reduce false positives, only pathways that contained between 10 and 200
1082 genes were used. Pathway size was determined by the total number of genes in the pathway that
1083 could also be mapped to the genes included in the GWAS dataset (actual pathway size may be
1084 larger).

1085

1086 SNPs were assigned to genes using the INQUISIT target prediction method described above for all
1087 SNPs with P-value < $5x10^{-2}$ (~1.25 million associations). This cutoff was chosen based on a threshold
1088 analysis that showed that 19 of the 20 pathway themes found using all SNP associations (~16 million)
1089 and a simple distance-based SNP-to-gene mapping method could be recovered using this smaller
1090 subset of associations. More stringent cutoffs resulted in fewer themes being covered (e.g. three
1091 themes found using SNPs with p-value < $5x10^{-6}$ or ~33K SNP associations). Gene significance was
1092 calculated by assigning the statistic of the most significant SNP among all SNPs assigned to a
1093 gene[68,69]. Since histone genes contained a high number of mapped SNPs, we selected representative
1094 SNP associations to avoid pathway enrichments based solely on the increased number of SNPs at
1095 these loci (i.e. chr6:27657944 for HIST1, chr1:149219841, for HIST2, chr1: 228517406 for HIST3,
1096 chr12: 14871747 for HIST4).

1097

1098 The gene set enrichment analysis (GSEA) algorithm as implemented in the GenGen package[69] was
1099 used to perform pathway analysis. Wang et al.[70] modified the original GSEA algorithm to work with
1100 GWAS datasets, using SNP significance and SNP-to-gene mapping instead of gene expression data.
1101 Briefly, the algorithm calculates an enrichment score (ES) for each pathway based on a weighted
1102 Kolmogorov-Smirnov statistic (refer to [70] for more details). Pathways that have most of their genes
1103 at the top of the ranked list of genes obtain higher ES values. Note that only the largest positive ES
1104 was considered as opposed to largest absolute ES (i.e. largest deviation from zero). This modification
1105 (recommended by the GenGen authors for GWAS analysis) was performed to include only pathways
1106 that are significantly affected between cases and controls and ignore those with significant negative
1107 ES values (this may happen if a pathway is significantly less altered than expected by chance). Only
1108 pathways containing greater than 10 genes with at least one of these genes with P-value < $5x10^{-8}$
1109 were retained as higher confidence for subsequent analysis. These pathways, together with the
1110 genes reaching the significance threshold, are listed in **Supplementary Table 21**.

1111

1112 The pathway analysis assigns an enrichment score (ES) value for each pathway. These values were
1113 normalized and p-values for each pathway were obtained by comparing them to null distributions
1114 for OncoArray and iCOGS data sets separately. The null distributions were computed by permuting
1115 case/control labels 1,000 times (keeping the number of cases and controls the same in each
1116 iteration) and recomputing all enrichment statistics. FDR values were computed using the statistics
1117 from the null distributions and all pathways with FDR < 0.05 in either OncoArray or iCOGS
1118 distributions were considered further. Pathway findings were further considered if they contained

1119 more than one significant gene and if they could be confirmed to be involved in breast cancer as
1120 reported in at least one of five published large-scale breast cancer GWAS[71-75] or reported elsewhere
1121 in the literature. Further, themes that were weakly associated with breast cancer (based on a
1122 literature search) were only included if they had a FDR < 0.05 and at least four novel genes (i.e. was
1123 not found among the genes from mapped themes containing pathways known to be involved in
1124 breast cancer) (Extended Data Fig. 2). Pathways related to "sensory perception of smell" were
1125 removed as there is no literature evidence for their involvement in breast cancer and because they
1126 contain genes close to each other on chromosome 6 which are frequently correlated.
1127
1128 An enrichment map was created using the Enrichment Map (EM) v 2.1.0 app[13] in Cytoscape v 3.3[76].
1129 Pathways nodes were laid out using a force directed layout and nodes with gene set overlap of over
1130 0.55 were connected by edges. Related pathway nodes were manually clustered and labelled as
1131 themes.
1132

1133 **References**

1134

1135 20 Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals
1136 identifies 15 new susceptibility loci for breast cancer. *Nature genetics* **47**, 373-380,
1137 doi:10.1038/ng.3242 (2015).
1138 21 O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of
1139 relatedness. *PLoS Genet* **10**, e1004234, doi:10.1371/journal.pgen.1004234 (2014).
1140 22 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation
1141 method for the next generation of genome-wide association studies. *PLoS Genet* **5**,
1142 e1000529, doi:10.1371/journal.pgen.1000529 (2009).
1143 23 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate
1144 genotype imputation in genome-wide association studies through pre-phasing. *Nature*
1145 *genetics* **44**, 955-959, doi:10.1038/ng.2354 (2012).
1146 24 Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype
1147 data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-834,
1148 doi:10.1002/gepi.20533 (2010).
1149 25 Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide
1150 association analysis of imputed data. *BMC Bioinformatics* **11**, 134, doi:10.1186/1471-2105-
1151 11-134 (2010).
1152 26 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide
1153 association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).
1154 27 Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than
1155 replication-based analysis for two-stage genome-wide association studies. *Nature genetics*
1156 **38**, 209-213, doi:10.1038/ng1706 (2006).
1157 28 Team, R. C. *R: A Language and Environment for Statistical Computing* <https://www.R-
1158 project.org> (2016).
1159 29 Consortium, E. P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**,
1160 e1001046, doi:10.1371/journal.pbio.1001046 (2011).
1161 30 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes.
1162 *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
1163 31 Udler, M. S., Tyrer, J. & Easton, D. F. Evaluating the power to discriminate between highly
1164 correlated SNPs in genetic association studies. *Genet Epidemiol* **34**, 463-468,
1165 doi:10.1002/gepi.20504 (2010).
1166 32 Wellcome Trust Case Control, C. *et al.* Bayesian refinement of association signals for 14 loci
1167 in 3 common diseases. *Nature genetics* **44**, 1294-1301, doi:10.1038/ng.2435 (2012).
1168 33 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals
1169 novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).
1170 34 Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations.
1171 *Bioinformatics* **28**, 1359-1367, doi:10.1093/bioinformatics/bts144 (2012).
1172 35 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human
1173 genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
1174 36 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or
1175 without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323
1176 (2011).
1177 37 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of
1178 focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41,
1179 doi:10.1186/gb-2011-12-4-r41 (2011).
1180 38 Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*
1181 **152**, 633-641, doi:10.1016/j.cell.2012.12.034 (2013).
1182 39 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
1183 *Bioinformatics* **28**, 1353-1358, doi:10.1093/bioinformatics/bts163 (2012).

1184 40 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-
1185 496, doi:10.1093/nar/gkh103 (2004).
1186 41 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of
1187 chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
1188 42 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1189 features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
1190 43 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE
1191 Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).
1192 44 Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage
1193 disequilibrium dictate levels of gene expression to confer susceptibility to common traits.
1194 *Genome Res* **24**, 1-13, doi:10.1101/gr.164079.113 (2014).
1195 45 He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human
1196 cells. *Proc Natl Acad Sci U S A* **111**, E2191-2199, doi:10.1073/pnas.1320308111 (2014).
1197 46 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature*
1198 **507**, 455-461, doi:10.1038/nature12787 (2014).
1199 47 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947,
1200 doi:10.1016/j.cell.2013.09.053 (2013).
1201 48 Joly Beauparlant, C. *et al.* metagene Profiles Analyses Reveal Regulatory Element's Factor-
1202 Specific Recruitment Patterns. *PLoS Comput Biol* **12**, e1004751,
1203 doi:10.1371/journal.pcbi.1004751 (2016).
1204 49 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
1205 genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).
1206 50 Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of
1207 amino acid substitutions using hidden Markov models. *Human mutation* **34**, 57-65,
1208 doi:10.1002/humu.22225 (2013).
1209 51 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes.
1210 *Genome Res* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).
1211 52 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations:
1212 application to cancer genomics. *Nucleic Acids Res* **39**, e118, doi:10.1093/nar/gkr407 (2011).
1213 53 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation
1214 prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362, doi:10.1038/nmeth.2890
1215 (2014).
1216 54 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat*
1217 *Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).
1218 55 Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of
1219 amino acid substitutions and indels. *PloS one* **7**, e46688, doi:10.1371/journal.pone.0046688
1220 (2012).
1221 56 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants
1222 on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081,
1223 doi:10.1038/nprot.2009.86 (2009).
1224 57 Desmet, F. O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing
1225 signals. *Nucleic Acids Res* **37**, e67, doi:10.1093/nar/gkp215 (2009).
1226 58 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with
1227 applications to RNA splicing signals. *J Comput Biol* **11**, 377-394,
1228 doi:10.1089/1066527041410418 (2004).
1229 59 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.
1230 *Nucleic Acids Res* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).
1231 60 Ghoussaini, M. *et al.* Evidence that breast cancer risk at the 2q35 locus is mediated through
1232 IGFBP5 regulation. *Nat Commun* **4**, 4999, doi:10.1038/ncomms5999 (2014).
1233 61 Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*
1234 **33**, D428-432, doi:10.1093/nar/gki072 (2005).

1235 62 Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-679,
1236 doi:10.1093/nar/gkn653 (2009).
1237 63 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology
1238 Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
1239 64 Romero, P. *et al.* Computational prediction of human metabolic pathways from the
1240 complete human genome. *Genome Biol* **6**, R2, doi:10.1186/gb-2004-6-1-r2 (2005).
1241 65 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
1242 interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550,
1243 doi:10.1073/pnas.0506580102 (2005).
1244 66 Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways.
1245 *Genome Biol* **11**, R3, doi:10.1186/gb-2010-11-1-r3 (2010).
1246 67 Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by
1247 function. *Genome Res* **13**, 2129-2141, doi:10.1101/gr.772403 (2003).
1248 68 Wang, L., Jia, P., Wolfinger, R. D., Chen, X. & Zhao, Z. Gene set analysis of genome-wide
1249 association studies: methodological issues and perspectives. *Genomics* **98**, 1-8,
1250 doi:10.1016/j.ygeno.2011.04.006 (2011).
1251 69 Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association
1252 studies. *Nat Rev Genet* **11**, 843-854, doi:10.1038/nrg2884 (2010).
1253 70 Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide
1254 association studies. *Am J Hum Genet* **81**, 1278-1283, doi:10.1086/522374 (2007).
1255 71 Mogushi, K. & Tanaka, H. PathAct: a novel method for pathway analysis using gene
1256 expression profiles. *Bioinformation* **9**, 394-400, doi:10.6026/97320630009394 (2013).
1257 72 Medina, I. *et al.* Gene set-based analysis of polymorphisms: finding pathways or biological
1258 processes associated to traits in genome-wide association studies. *Nucleic Acids Res* **37**,
1259 W340-344, doi:10.1093/nar/gkp481 (2009).
1260 73 Lee, Y. H., Kim, J. H. & Song, G. G. Genome-wide pathway analysis of breast cancer. *Tumour*
1261 *Biol* **35**, 7699-7705, doi:10.1007/s13277-014-2027-5 (2014).
1262 74 Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. dmGWAS: dense module searching for
1263 genome-wide association studies in protein-protein interaction networks. *Bioinformatics* **27**,
1264 95-102, doi:10.1093/bioinformatics/btq615 (2011).
1265 75 Braun, R. & Buetow, K. Pathways of distinction analysis: a new technique for multi-SNP
1266 analysis of GWAS data. *PLoS Genet* **7**, e1002101, doi:10.1371/journal.pgen.1002101 (2011).
1267 76 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular
1268 interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

1269

1270 **Supplementary Information** is linked to the online version of the paper at
1271 www.nature.com/nature.
1272
1273 **Acknowledgements**
1274
1275 The authors wish to thank all the individuals who took part in these studies and all the
1276 researchers, clinicians, technicians and administrative staff who have enabled this work to be
1277 carried out.

1278 Genotyping of the OncoArray was principally funded from three sources: the PERSPECTIVE
1279 project, funded from the Government of Canada through Genome Canada and the Canadian
1280 Institutes of Health Research, the *Ministère de l'Économie, de la Science et de l'Innovation du*
1281 *Québec* through Genome Québec, and the Quebec Breast Cancer Foundation; the NCI Genetic
1282 Associations and Mechanisms in Oncology (GAME-ON) initiative and Discovery, Biology and
1283 Risk of Inherited Variants in Breast Cancer (DRIVE) project (NIH Grants U19 CA148065 and
1284 X01HG007492); and Cancer Research UK (C1287/A10118 and. C1287/A16563). **BCAC** is funded
1285 by Cancer Research UK [C1287/A16563], by the European Community's Seventh Framework
1286 Programme under grant agreement 223175 (HEALTH-F2-2009-223175) (COGS) and by the
1287 European Union's Horizon 2020 Research and Innovation Programme under grant agreements
1288 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the iCOGS array was funded by the
1289 European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the
1290 Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer"
1291 program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec –
1292 grant # PSR-SIIRI-701. Combining the GWAS data was supported in part by The National
1293 Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE, part
1294 of the GAME-ON initiative). For a full description of funding and acknowledgments, see
1295 **Supplementary Note**.

1296

1297 **Author Contributions**

1298 Writing Group: K.Michailidou, S.Lindström, J.Beesley, S.Hui, S.Kar, P.Soucy, S.L.E., G.D.B., G.C-T.,
1299 J.Simard, P.K., D.F.E.

1300 Conceived OncoArray and obtained financial support: C.I.A., J.Simard, P.K., D.F.E.

1301 Designed OncoArray: J.D., E.D., A.Lee, Z.W., A.C.A., S.J.C., P.K., D.F.E.

1302 Led COGS project: P.Hall.

1303 Led DRIVE project: D.J.H.

1304 Led PERSPECTIVE project: J.Simard.

1305 Led working groups of BCAC: A.C.A., I.L.A., P.D.P.P., J.Chang-Claude, R.L.M., M.G-C., M.K.S.,
1306 A.M.D.

1307 Data management: J.D., M.K.B., Q.Wang, R.Keeman, U.E., S.B., J.Chang-Claude, M.K.S.

1308 Bioinformatics analysis: J.D., J.Beesley, A.Lemaçon, P.Soucy, J.A., M.Ghoussaini, J.Carroll, A.D.,
1309 A.E. McC R, S.R.L.

1310 Statistical analysis: K.Michailidou, S.Lindström, S.Hui, S.Kar, A.Rostamianfar, J.T., X.C., L.Fachal,
1311 X.J., H.Finucane, G.D.B., P.K., D.F.E.

1312 Functional analysis: D.G., X.C., J.Beesley, J.D.F., K.McCue, S.L.E., G.C-T.

1313 OncoArray Genotyping: M.A., F.B., C.Baynes, D.M.C., J.M.C., K.F.D., N.Hamel, B.H., K.J., C.L.,
1314 J.Meyer, E.P., J.R., G.S., D.C.T., D.V.D.B., D.V., J.V., L.X., B.Z., A.M.D.

1315 Provided DNA samples and/or phenotypic data: M.A.A., K.A., H.A-C., N.N.A., V.A., K.J.A., B.A.,
1316 P.L.A., M.Barrdahl, M.W.B., J.Benitez, M.Bermisheva, L.Bernstein, C.Blomqvist, N.V.B., S.E.B.,
1317 B.Bonanni, A-L.B-D., J.S.B., H.Brauch, P.Brennan, H.Brenner, L.Brinton, P.Broberg, I.W.B., A.B.,
1318 A.B-W., S.Y.B., T.B., B.Burwinkel, K.B., H.Cai, Q.C., T.C., F.C., A.Carracedo, B.D.C., J.EstebanC.,
1319 T.L.C., T-Y.D.C, K.S.C., J-Y.Choi, H.Christiansen, C.L.C., M.C., E.C-D., S.C., A.Cox, D.C., S.S.C., K.C.,
1320 M.B.D., H.D., P.D., T.D., I.d.S.S., M.Dumont, L.D., M.Dwek, D.M.E., A.B.E., A.H.E., C.Ellberg,
1321 M.Elvira, C.Engel, M.Eriksson, P.A.F., J.F., D.F-J., O.F., H.Flyger, L.Fritschi, V.Gaborieau, M.G.,
1322 M.G-D., Y-T.G., S.M.G., J.A.G-S., M.M.G., V.Georgoulias, G.G.G., G.G., M.S.G., D.E.G., A.G-N.,
1323 G.I.G., M.Grip, J.G., A.G., P.G., L.H., E.H., C.A.H., N.Håkansson, U.H., S.Hankinson, P.Harrington,
1324 S.N.H., J.M.H., M.H., A.Hein, J.H., P.Hillemanns, D.N.H., A.Hollestelle, M.J.H., R.N.H., J.L.H., M-
1325 F.H., C-N.H., G.H., K.H., J.I., H.Ito, M.I., H., A.J., W.J., E.M.J., N.J., M.J., A.J-V., R.Kaaks, M.K., K.K.,
1326 D.K., Y.K., M.J.K., S.Khan, E.K., J.I.K., S-W.K., J.A.K., V-M.K., V.N.K., U.K., A.K., D.L., L.L., C.N.L., E.L.,
1327 J.W.L., M.L., F.L., J.Li, J.Lilyquist, A.Lindblom, J.Lissowska, W-Y.L., S.Loibl, J.Long,

1328 A.Lophatananon, J.Lubinski, M.P.L., E.S.K.M., R.J.M., T.M., E.M., I.M., A.Mannermaa,
1329 S.Manoukian, J.E.M., S.Margolin, S.Mariapun, M.E.M., K.Matsuo, D.M., J.McKay, C.McLean, H.M-
1330 H., A.Meindl, P.M., U.M., H.M., N.M., N.A.M.T., K.Muir, A.M.M., C.Mulot, S.L.N., H.N., P.N.,
1331 S.F.N., D-Y.N., B.G.N., A.N., O.I.O., J.E.O., H.O., C.O., N.O., V.P., S.K.P., T-W.P-S., R.Peake, J.I.A.P.,
1332 P.P., J.P., K-A.P., M.P., D.P-K., R.Prentice, N.P., D.P., K.P., B.R., P.R., N.R., G.R., H.S.R., V.R.,
1333 A.Romero, K.R., T.R., A.Rudolph, M.R., E.J.Th.R., E.S., D.P.S., S.Sangrajrang, E.J.S., D.F.S., R.K.S.,
1334 A.Schneeweiss, M.J.Schoemaker, F.S., P.Schürmann, C.Scott, R.J.S., S.Seal, C.Seynaeve, M.S.,
1335 P.Sharma, C-Y.S., M.E.S., M.J.Shrubsole, X-O.S., A.Smeets, C.Sohn, M.C.S., J.J.S., C.Stegmaier, S.S-
1336 B., J.Stone, D.O.S., H.S., A.Swerdlow, R.T., J.A.T., M.T., S.H.T., M.B.T., S.Thanasitthichai, K.T.,
1337 R.A.E.M.T., I.T., D.T., T.T., C-C.T., S.Tsugane, H-U.U., M.U., C.V., C.J.v.A., A.M.W.v.d.O., L.v.d.K.,
1338 R.B.v.d.L., Q.Waisfisz, S.W-G., C.R.W., C.W., A.S.W., H.W., W.W., R.W., A.W., A.H.W., T.Y., X.R.Y.,
1339 C.Y., K-Y.Y., J-C.Y., W.Z., Y.Z., A.Z., E.Z., ABCTB.I., kConFab/AOCS.I., NCBS.C., A.C.A., I.L.A., F.J.C.,
1340 P.D.P.P., J.Chang-Claude, P.Hall, D.J.H., R.L.M., M.G-C., M.K.S., G.D.B., J.Simard, P.K., D.F.E.
1341 All authors read and approved the final version of the manuscript.

1342 **Author Information.** A subset of the data that support the findings of this study is publically
1343 available via dbGaP (www.ncbi.nlm.nih.gov/gap; accession number phs001265.v1.p1). The
1344 complete dataset will not be made publicly available due to restraints imposed by the ethics
1345 committees of individual studies; requests for data can be made to the corresponding author or
1346 the Data Access Coordination Committee (DACCs) of BCAC
1347 (http://bcac.ccge.medschl.cam.ac.uk/): BCAC DACC approval is required to access data from
1348 studies ABCFS, ABCS, ABCTB, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UT, BCINIS, BSUCH,
1349 CBCS, CECILE, CGPS, CTS, DIETCOMPLYF, ESTHER, GC-HBOC, GENICA, GEPARSIXTO, GESBC,
1350 HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KBCP, LMBC, MABCS, MARIE, MBCSG, MCBCS,
1351 MISS, MMHS, MTLGEBCS, NC-BCFR, OFBCR, ORIGO, pKARMA, POSH, PREFACE, RBCS, SKKDKFZS,
1352 SUCCESSB, SUCCESSC, SZBCS, TNBCC, UCIBCS, UKBGS and UKOPS (see **Supplementary Table 1**).

1353 Summary results for all variants are available at http://bcac.ccge.medschl.cam.ac.uk/. Requests
1354 for further data should be made through the BCAC Data Access Co-ordinating Committee
1355 (http://bcac.ccge.medschl.cam.ac.uk/). Reprints and permissions information is available
1356 through www.nature.com/reprints. The authors confirm that they have no competing financial
1357 interests. Correspondence should be addressed to D.F.E. (dfe20@medschl.cam.ac.uk).

1358

1359 **Extended Data Table 1: INQUISIT, DEPICT, and nearest gene as predictors of driver status.**
1360 Scores converted into levels for analysis. For INQUISIT: level 1 (coding score of 2 OR promoter
1361 score of 3 or 4 OR distal score > 4), level 2 (coding score of 1 OR promoter of 1 or 2 OR distal
1362 score of 1, 2, 3, or 4), level 3 (coding/promoter/distal scores > 0 but < 1), and level 4 (not
1363 predicted to be a target gene by INQUISIT). For DEPICT: level 1 (DEPICT predicted target gene at
1364 $P \leq 0.05$), level 2 (DEPICT predicted target gene but with $P > 0.05$), level 3 (not predicted to be a
1365 target gene by DEPICT).
1366
1367 **Extended Data Figure 1: Global mapping of biofeatures across novel loci associated with**
1368 **overall breast cancer risk.** The overlaps between potential genomic predictors in relevant breast
1369 cell lines and candidate causal risk variants (CRVs) within each locus. On the x-axis, each column
1370 represents a CRV (see **Online Methods**). The most significant SNPs are identified in each region.
1371 On the y-axis, biofeatures are grouped into five functional categories: genomic structure (red),
1372 enhancer marks (dark green), histone marks (blue), open chromatin marks (dark blue) and
1373 transcription factor binding sites (dark violet). Colored elements indicate SNPs for which the
1374 feature is present. For data sources, see **Online Methods** ("In-Silico Analysis of CRVs").
1375
1376 **Extended Data Figure 2: Pathway enrichment map for susceptibility loci based on summary**
1377 **association statistics.** Each circle (node) represents a pathway (gene set), coloured by
1378 enrichment score (ES) where redder nodes indicate lower FDRs. Larger nodes indicate pathways
1379 with more genes. Green lines connect pathways with overlapping genes (minimum overlap
1380 0.55). Pathways are grouped by similarity and organized into major themes (large labelled
1381 circles).
1382
1383 **Extended Data Figure 3. Heatmap showing patterns of cell type-specific enrichments for**
1384 **breast tissue across three histone marks (H3K4me1, H3K4me3 and H3K9ac) for breast cancer**
1385 **overall, ER-positive breast cancer and ER-negative breast cancer as well as 16 other traits.**
1386
1387 **Extended Data Figure 4: Heatmap showing patterns of cell type-specific enrichments for**
1388 **histone mark H3K27ac in breast cancer overall, ER+ and ER- breast cancer as well as 16**
1389 **other traits.**
1390

Extended Data Figure 5: Heatmap showing patterns of cell type-specific enrichments for histone mark H3K4me1 in breast cancer overall, ER+ and ER- breast cancer as well as 16 other traits.

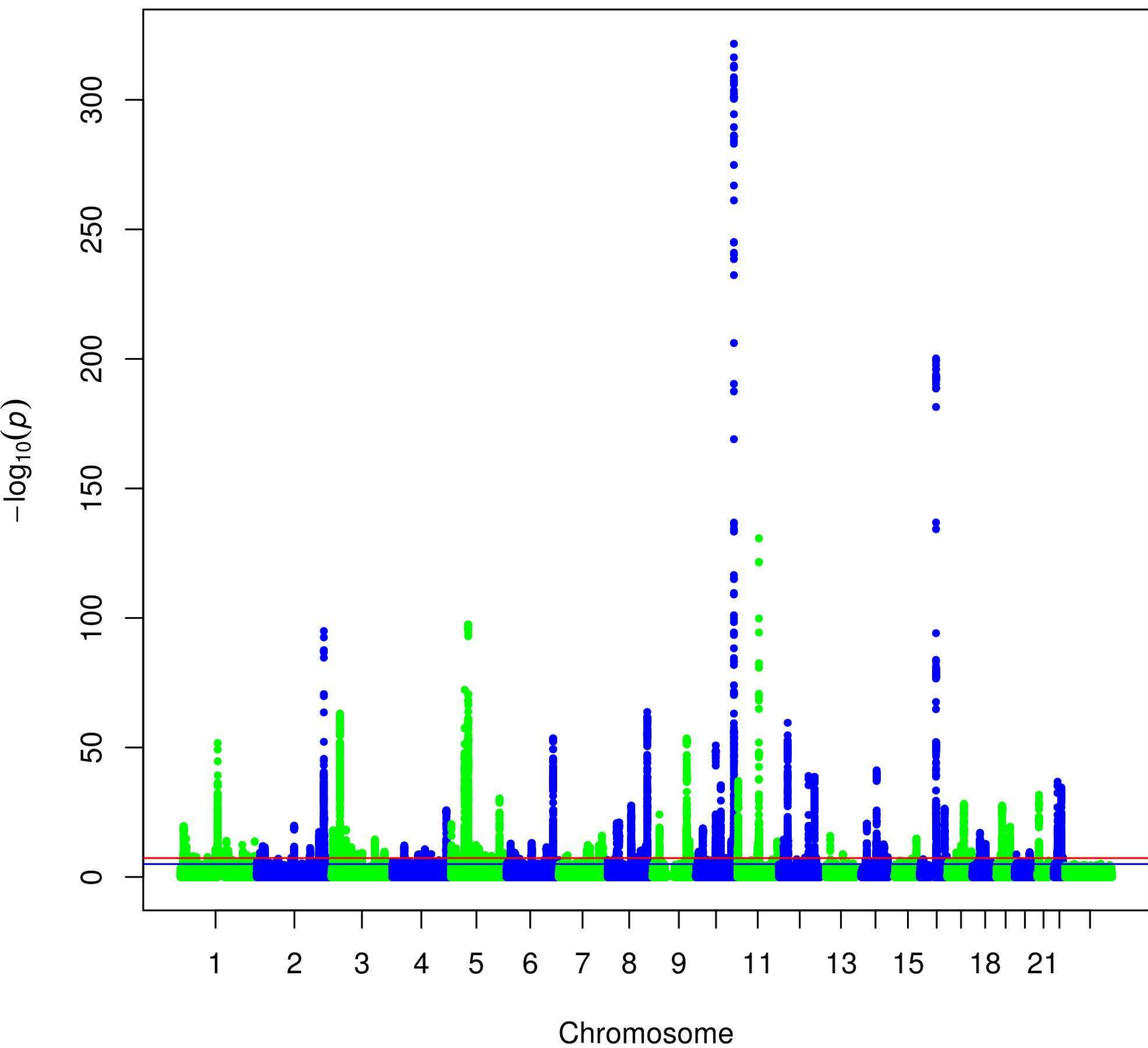
Extended Data Figure 6: Heatmap showing patterns of cell type-specific enrichments for histone mark H3K4me3 in breast cancer overall, ER+ and ER- breast cancer as well as 16 other traits.

Extended Data Figure 7: Heatmap showing patterns of cell type-specific enrichments for histone mark H3K9ac in breast cancer overall, ER-positive and ER-negative breast cancer as well as 16 other traits.

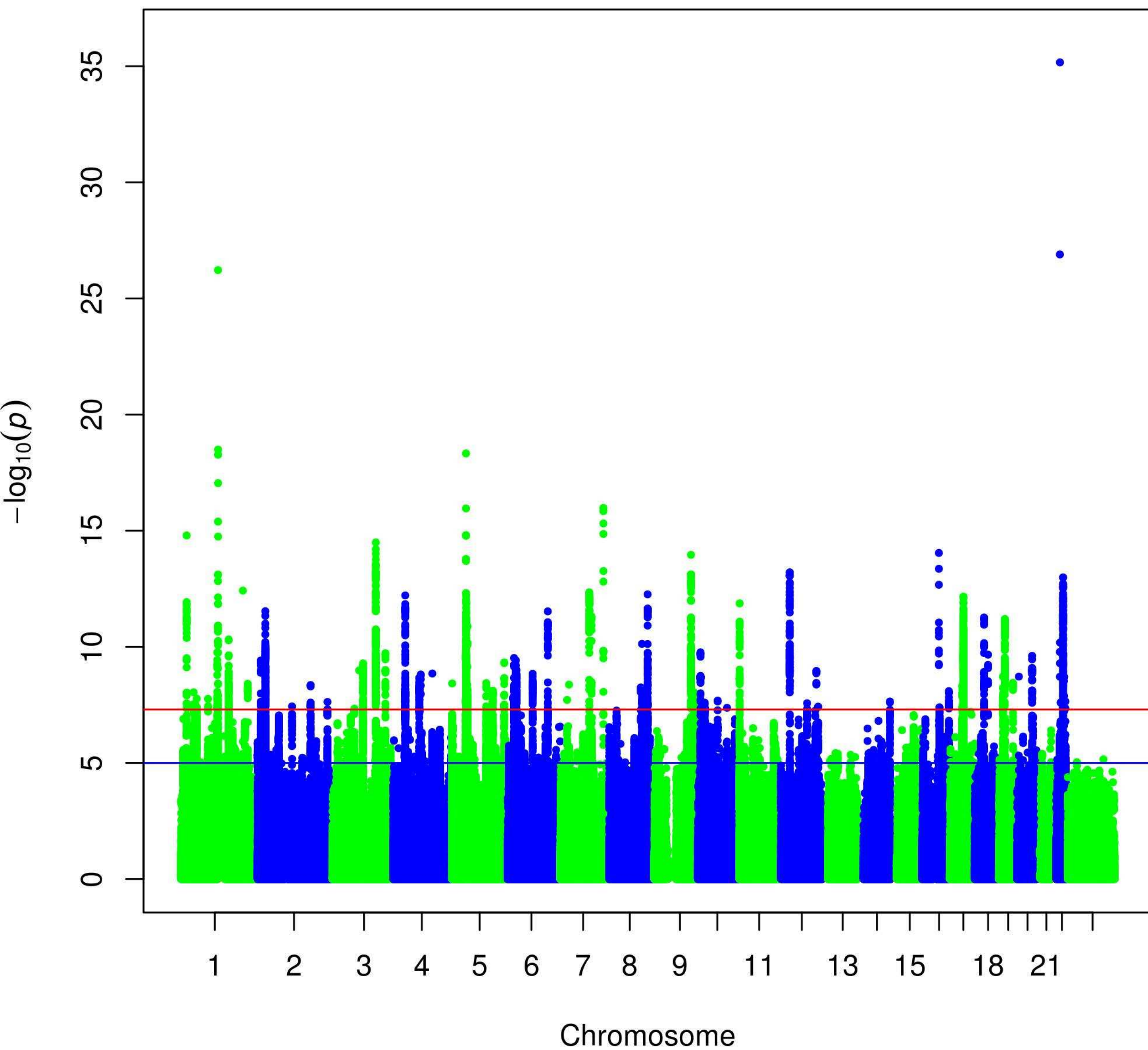
Extended Data Figure 8: Functional assessment of regulatory variants at 1p36, 11p15 and 1p34 risk loci. a, The *KLHDC7A* or **b**, *PIDD1* promoter regions containing the reference (prom-Ref) or risk alleles (prom-Hap), were cloned upstream of the pGL3 luciferase reporter gene. MCF7 or Bre-80 cells were transfected with constructs and assayed for luciferase activity after 24 h. Error bars denote 95% CI (n=3). *P*-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (**P*<0.05, ***P*<0.01, ****P*<0.001). **c**, 3C assays. A physical map of the region interrogated by 3C is shown first. Grey boxes depict the putative regulatory elements (PREs), blue vertical lines indicate the risk-associated SNPs and black dotted line represents chromatin looping. The graphs represent three independent 3C interaction profiles. 3C libraries were generated with *Eco*RI, grey vertical boxes indicate the interacting restriction fragment (containing PRE1 and PRE2). Error bars denote SD. **d**, PRE1 or PRE2 containing the reference (PRE-ref) or risk (PRE-Hap) haplotypes were cloned downstream of a *CITED4* promoter-driven luciferase construct (*CITED4* prom). MCF7 or Bre-80 cells were transfected with constructs and assayed for luciferase activity after 24 h. Error bars denote 95% CI (n=3). P-values were determined by two-way ANOVA followed by Dunnett's multiple comparisons test (**P<0.01, ***P<0.001).

Extended Data Figure 9: Functional assessment of regulatory variants at the 7q22 risk locus. a-e, 3C assays. A physical map of the region interrogated by 3C is shown first. Grey horizontal boxes depict the putative regulatory elements (PREs), blue vertical lines indicate the risk-associated SNPs and black dotted line represents chromatin looping. The graphs represent three independent 3C interaction profiles between the **a,** *CUX1*, **b**, **d,** *PRKRIP1* or **c**, **e,** *RASA4* promoter

1425 regions and PREs. 3C libraries were generated with *Eco*RI, grey vertical boxes indicate the
1426 interacting restriction fragment (containing PRE1 and/or PRE2). Error bars denote SD. **f, g,** Allele-
1427 specific 3C. 3C followed by Sanger sequencing for the **f**, *PRKRIP1*-PRE2 or **g**, *RASA4*-PRE1 or -
1428 PRE2 in heterozygous MDA-MB-231 breast cancer cells.
1429
1430



−log10(p)
0
50
100
150
200
250
300
1
2
3
4
5
6
7
8
9
11
13
15
18
21
Chromosome



35
30
25
20
15
10
5
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
11
13
15
18
21
Chromosome