

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



## LSHTM Research Online

Wu, L; Shi, W; Long, J; Guo, X; Michailidou, K; Beesley, J; Bolla, MK; Shu, XO; Lu, Y; Cai, Q; +202 more... Al-Ejeh, F; Rozali, E; Wang, Q; Dennis, J; Li, B; Zeng, C; Feng, H; Gusev, A; Barfield, RT; Andrulis, IL; Anton-Culver, H; Arndt, V; Aronson, KJ; Auer, PL; Barrdahl, M; Baynes, C; Beckmann, MW; Benitez, J; Bermisheva, M; Blomqvist, C; Bogdanova, NV; Bojesen, SE; Brauch, H; Brenner, H; Brinton, L; Broberg, P; Brucker, SY; Burwinkel, B; Caldés, T; Canzian, F; Carter, BD; Castelao, JE; Chang-Claude, J; Chen, X; Cheng, TD; Christiansen, H; Clarke, CL; NBCS Collaborators; Collée, M; Cornelissen, S; Couch, FJ; Cox, D; Cox, A; Cross, SS; Cunningham, JM; Czene, K; Daly, MB; Devilee, P; Doheny, KF; Dörk, T; Dos-Santos-Silva, I; Dumont, M; Dwek, M; Eccles, DM; Eilber, U; Eliassen, AH; Engel, C; Eriksson, M; Fachal, L; Fasching, PA; Figueroa, J; Flesch-Janys, D; Fletcher, O; Flyger, H; Fritschi, L; Gabrielson, M; Gago-Dominguez, M; Gapstur, SM; García-Closas, M; Gaudet, MM; Ghoussaini, M; Giles, GG; Goldberg, MS; Goldgar, DE; González-Neira, A; Guénel, P; Hahnen, E; Haiman, CA; Håkansson, N; Hall, P; Hallberg, E; Hamann, U; Harrington, P; Hein, A; Hicks, B; Hillemanns, P; Hollestelle, A; Hoover, RN; Hopper, JL; Huang, G; Humphreys, K; Hunter, DJ; Jakubowska, A; Janni, W; John, EM; Johnson, N; Jones, K; Jones, ME; Jung, A; Kaaks, R; Kerin, MJ; Khusnutdinova, E; Kosma, VM; Kristensen, VN; Lambrechts, D; Le Marchand, L; Li, J; Lindström, S; Lissowska, J; Lo, WY; Loibl, S; Lubinski, J; Luccarini, C; Lux, MP; MacInnis, RJ; Maishman, T; Kostovska, IM; Mannermaa, A; Manson, JE; Margolin, S; Mavroudis, D; Meijers-Heijboer, H; Meindl, A; Menon, U; Meyer, J; Mulligan, AM; Neuhausen, SL; Nevanlinna, H; Neven, P; Nielsen, SF; Nordestgaard, BG; Olopade, OI; Olson, JE; Olsson, H; Peterlongo, P; Peto, J; Plaseska-Karanfilska, D; Prentice, R; Presneau, N; Pylkäs, K; Rack, B; Radice, P; Rahman, N; Rennert, G; Rennert, HS; Rhenius, V; Romero, A; Romm, J; Rudolph, A; Saloustros, E; Sandler, DP; Sawyer, EJ; Schmidt, MK; Schmutzler, RK; Schneeweiss, A; Scott, RJ; Scott, CG; Seal, S; Shah, M; Shrubsole, MJ; Smeets, A; Southey, MC; Spinelli, JJ; Stone, J; Surowy, H; Swerdlow, AJ; Tamimi, RM; Tapper, W; Taylor, JA; Terry, MB; Tessier, DC; Thomas, A; Thöne, K; Tollenaar, RAEM; Torres, D; Truong, T; Untch, M; Vachon, C; Van Den Berg, D; Vincent, D; Waisfisz, Q; Weinberg, CR; Wendt, C; Whittemore, AS; Wildiers, H; Willett, WC; Winqvist, R; Wolk, A; Xia, L; Yang, XR; Ziogas, A; Ziv, E; kConFab/AOCS Investigators; Dunning, AM; Pharoah, PDP; Simard, J; Milne, RL; Edwards, SL; Kraft, P; Easton, DF; Chenevix-Trench, G; Zheng, W; (2018) A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature genetics*. ISSN 1061-4036 DOI: <https://doi.org/10.1038/s41588-018-0132-x>

Downloaded from: <http://researchonline.lshtm.ac.uk/4648310/>

DOI: <https://doi.org/10.1038/s41588-018-0132-x>

**Usage Guidelines:**

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact [researchonline@lshtm.ac.uk](mailto:researchonline@lshtm.ac.uk).

Available under license: <http://creativecommons.org/licenses/by-nc-nd/2.5/>

<https://researchonline.lshtm.ac.uk>

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

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

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

51  
 52 **\*Corresponding Authors:** Wei Zheng, MD, PhD, Division of Epidemiology, Department of  
 53 Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt  
 54 University Medical Center, 2525 West End Ave, Suite 800, Nashville, Tennessee, 37203, USA.  
 55 Email: [wei.zheng@vanderbilt.edu](mailto:wei.zheng@vanderbilt.edu) and Georgia Chenevix-Trench, PhD, Cancer Division, QIMR  
 56 Berghofer Medical Research Institute, 300 Herston Road, Herston 4006, Australia. Email:  
 57 [Georgia.Trench@qimrberghofer.edu.au](mailto:Georgia.Trench@qimrberghofer.edu.au)

58  
 59

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

- 61
- 62 1. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center,  
 63 Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville,  
 64 TN, USA.
  - 65 2. Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, Australia.
  - 66 3. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary  
 67 Care, University of Cambridge, Cambridge, UK.
  - 68 4. Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of  
 69 Neurology and Genetics, Nicosia, Cyprus.
  - 70 5. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of  
 71 Public Health, Boston, MA, USA.
  - 72 6. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA,  
 73 USA.
  - 74 7. Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of  
 75 Mount Sinai Hospital, Toronto, ON, Canada.
  - 76 8. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada.
  - 77 9. Department of Epidemiology, University of California Irvine, Irvine, CA, USA.
  - 78 10. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center  
 79 (DKFZ), Heidelberg, Germany.
  - 80 11. Department of Public Health Sciences, and Cancer Research Institute, Queen's  
 81 University, Kingston, ON, Canada.
  - 82 12. Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA,  
 83 USA.
  - 84 13. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI,  
 85 USA.
  - 86 14. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg,  
 87 Germany.
  - 88 15. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of  
 89 Cambridge, Cambridge, UK.
  - 90 16. Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-  
 91 Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-  
 92 EMN, Erlangen, Germany.

- 93 17. Human Cancer Genetics Program, Spanish National Cancer Research Centre, Madrid,  
94 Spain.
- 95 18. Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain.
- 96 19. Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of  
97 Sciences, Ufa, Russia.
- 98 20. Gynaecology Research Unit, Hannover Medical School, Hannover, Germany.
- 99 21. Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki,  
100 Finland.
- 101 22. Department of Radiation Oncology, Hannover Medical School, Hannover, Germany.
- 102 23. N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk,  
103 Belarus.
- 104 24. Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen  
105 University Hospital, Herlev, Denmark.
- 106 25. Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen  
107 University Hospital, Herlev, Denmark.
- 108 26. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen,  
109 Denmark.
- 110 27. Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany.
- 111 28. University of Tübingen, Tübingen, Germany.
- 112 29. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ),  
113 Heidelberg, Germany.
- 114 30. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National  
115 Center for Tumor Diseases (NCT), Heidelberg, Germany.
- 116 31. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville,  
117 MD, USA.
- 118 32. Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden.
- 119 33. Department of Gynecology and Obstetrics, University of Tübingen, Tübingen, Germany.
- 120 34. Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg,  
121 Germany.
- 122 35. Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ),  
123 Heidelberg, Germany.
- 124 36. Medical Oncology Department, CIBERONC Hospital Clínico San Carlos, Madrid, Spain.
- 125 37. Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg,  
126 Germany.
- 127 38. Epidemiology Research Program, American Cancer Society, Atlanta, GA, USA.
- 128 39. Oncology and Genetics Unit, Instituto de Investigacion Biomedica (IBI) Orense-  
129 Pontevedra-Vigo, Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain.
- 130 40. University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-  
131 Eppendorf, Hamburg, Germany.
- 132 41. Department of Epidemiology, University of Florida, Gainesville, FL, USA.
- 133 42. Westmead Institute for Medical Research, University of Sydney, Sydney, Australia.
- 134 43. Department of Oncology, Haukeland University Hospital, Bergen, Norway.
- 135 44. National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital  
136 Radiumhospitalet, Oslo, Norway.
- 137 45. Oslo University Hospital, Oslo, Norway.

- 138 46. Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The  
139 Netherlands.
- 140 47. Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van  
141 Leeuwenhoek Hospital, Amsterdam, The Netherlands.
- 142 48. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA.
- 143 49. Department of Epidemiology and Biostatistics, School of Public Health, Imperial College  
144 London, London, UK.
- 145 50. INSERM U1052, Cancer Research Center of Lyon, Lyon, France.
- 146 51. Sheffield Institute for Nucleic Acids, Department of Oncology and Metabolism,  
147 University of Sheffield, Sheffield, UK.
- 148 52. Academic Unit of Pathology, Department of Neuroscience, University of Sheffield,  
149 Sheffield, UK.
- 150 53. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm,  
151 Sweden.
- 152 54. Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA.
- 153 55. Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.
- 154 56. Department of Human Genetics, Leiden University Medical Center, Leiden, The  
155 Netherlands.
- 156 57. Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns  
157 Hopkins University School of Medicine, Baltimore, MD, USA.
- 158 58. Department of Non-Communicable Disease Epidemiology, London School of Hygiene  
159 and Tropical Medicine, London, UK.
- 160 59. Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval  
161 University, Québec City, QC, Canada.
- 162 60. Department of Biomedical Sciences, Faculty of Science and Technology, University of  
163 Westminster, London, UK.
- 164 61. Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton,  
165 Southampton, UK.
- 166 62. Channing Division of Network Medicine, Department of Medicine, Brigham and  
167 Women's Hospital, Harvard Medical School, Boston, MA, USA.
- 168 63. Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig,  
169 Leipzig, Germany.
- 170 64. David Geffen School of Medicine, Department of Medicine Division of Hematology and  
171 Oncology, University of California at Los Angeles, Los Angeles, CA, USA.
- 172 65. Usher Institute of Population Health Sciences and Informatics, The University of  
173 Edinburgh Medical School, Edinburgh, UK.
- 174 66. Institute for Medical Biometrics and Epidemiology, University Medical Center Hamburg-  
175 Eppendorf, Hamburg, Germany.
- 176 67. Department of Cancer Epidemiology, Clinical Cancer Registry, University Medical  
177 Center Hamburg-Eppendorf, Hamburg, Germany.
- 178 68. The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research,  
179 London, UK.
- 180 69. Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University  
181 Hospital, Herlev, Denmark.
- 182 70. School of Public Health, Curtin University, Perth, Australia.

- 183 71. Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de  
184 Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario  
185 Universitario de Santiago, SERGAS, Santiago De Compostela, Spain.
- 186 72. Moores Cancer Center, University of California San Diego, La Jolla, CA, USA.
- 187 73. Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne,  
188 Australia.
- 189 74. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global  
190 Health, The University of Melbourne, Melbourne, Australia.
- 191 75. Department of Medicine, McGill University, Montréal, QC, Canada.
- 192 76. Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University,  
193 Montréal, QC, Canada.
- 194 77. Department of Dermatology, Huntsman Cancer Institute, University of Utah School of  
195 Medicine, Salt Lake City, UT, USA.
- 196 78. Cancer & Environment Group, Center for Research in Epidemiology and Population  
197 Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif,  
198 France.
- 199 79. Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne,  
200 Cologne, Germany.
- 201 80. Center for Integrated Oncology (CIO), University Hospital of Cologne, Cologne,  
202 Germany.
- 203 81. Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne,  
204 Germany.
- 205 82. Department of Preventive Medicine, Keck School of Medicine, University of Southern  
206 California, Los Angeles, CA, USA.
- 207 83. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
- 208 84. Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.
- 209 85. Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ),  
210 Heidelberg, Germany.
- 211 86. Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National  
212 Laboratory for Cancer Research, Frederick, MD, USA.
- 213 87. Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute,  
214 Rotterdam, The Netherlands.
- 215 88. Department of Genetics and Pathology, Pomeranian Medical University, Szczecin,  
216 Poland.
- 217 89. Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm, Germany.
- 218 90. Department of Epidemiology, Cancer Prevention Institute of California, Fremont, CA,  
219 USA.
- 220 91. Department of Health Research and Policy - Epidemiology, Stanford University School  
221 of Medicine, Stanford, CA, USA.
- 222 92. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA.
- 223 93. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK.
- 224 94. School of Medicine, National University of Ireland, Galway, Ireland.
- 225 95. Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa,  
226 Russia.
- 227 96. Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland.

- 228 97. Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern  
229 Finland, Kuopio, Finland.
- 230 98. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio,  
231 Finland.
- 232 99. Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital  
233 Radiumhospitalet, Oslo, Norway.
- 234 100. Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway.
- 235 101. Department of Clinical Molecular Biology, Oslo University Hospital, University of Oslo,  
236 Oslo, Norway.
- 237 102. VIB KULeuven Center for Cancer Biology, VIB, Leuven, Belgium.
- 238 103. Laboratory for Translational Genetics, Department of Human Genetics, KU Leuven,  
239 Leuven, Belgium.
- 240 104. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA.
- 241 105. Department of Epidemiology, University of Washington School of Public Health, Seattle,  
242 WA, USA.
- 243 106. Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Institute -  
244 Oncology Center, Warsaw, Poland.
- 245 107. German Breast Group, GmbH, Neu Isenburg, Germany.
- 246 108. Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton,  
247 Southampton, UK.
- 248 109. Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov",  
249 Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia.
- 250 110. Department of Medicine, Brigham and Women's Hospital, Harvard Medical School,  
251 Boston, MA, USA.
- 252 111. Department of Oncology - Pathology, Karolinska Institutet, Stockholm, Sweden.
- 253 112. Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece.
- 254 113. Division of Gynaecology and Obstetrics, Technische Universität München, Munich,  
255 Germany.
- 256 114. Gynaecological Cancer Research Centre, Women's Cancer, Institute for Women's Health,  
257 University College London, London, UK.
- 258 115. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto,  
259 ON, Canada.
- 260 116. Laboratory Medicine Program, University Health Network, Toronto, ON, Canada.
- 261 117. Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte,  
262 CA, USA.
- 263 118. Department of Obstetrics and Gynecology, Helsinki University Hospital, University of  
264 Helsinki, Helsinki, Finland.
- 265 119. Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer  
266 Institute, University Hospitals Leuven, Leuven, Belgium.
- 267 120. Center for Clinical Cancer Genetics and Global Health, The University of Chicago,  
268 Chicago, IL, USA.
- 269 121. IFOM, The FIRC (Italian Foundation for Cancer Research) Institute of Molecular  
270 Oncology, Milan, Italy.
- 271 122. Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine  
272 Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland.



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

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

330  
331  
332  
333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349 **Abstract:**

350 Breast cancer risk variants identified in genome-wide association studies explain only a small  
351 fraction of familial relative risk, and genes responsible for these associations remain largely  
352 unknown. To identify novel risk loci and likely causal genes, we performed a transcriptome-wide  
353 association study evaluating associations of genetically predicted gene expression with breast  
354 cancer risk in 122,977 cases and 105,974 controls of European ancestry. We used data from 67  
355 subjects included in the Genotype-Tissue Expression Project to establish genetic models to  
356 predict gene expression in breast tissue and evaluated model performance using data from 86  
357 subjects included in The Cancer Genome Atlas. Of the 8,597 genes evaluated, significant  
358 associations were identified for 48 at a Bonferroni-corrected threshold of  $P < 5.82 \times 10^{-6}$ ,  
359 including 14 genes at loci not yet reported for breast cancer risk. We silenced 13 genes and  
360 showed an effect for 11 on cell proliferation and/or colony forming efficiency. Our study  
361 provides new insights into breast cancer genetics and biology.

362

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

371  
372 A large proportion of disease-associated risk variants identified by GWAS are located in non-  
373 protein coding or intergenic regions and are not in linkage disequilibrium (LD) with any  
374 nonsynonymous coding single nucleotide polymorphisms (SNPs)<sup>14</sup>. Many of these susceptibility  
375 variants are located in gene regulatory elements<sup>15,16</sup>, and it has therefore been hypothesized that  
376 most of the GWAS-identified associations may be driven by the regulatory function of risk  
377 variants on the expression levels of nearby genes. For breast cancer, recent studies have shown  
378 that GWAS-identified associations at 1p34, 1p36, 2q35, 5p12, 5p15.33, 5q11.2, 5q14, 6q25,  
379 7q22, 9q31.2, 10q21.3, 10q26.13, 11p15, 11q13.3, 15q26.1, 19p13 and 19q13.31 are likely due  
380 to the effect of risk variants at these loci on regulating the expression of either nearby or more  
381 distal genes: CITED4, KLHDC7A, IGFBP5, FGF10/MRPS30, TERT, MAP3K1, ATP6AP1L,  
382 RMND1, RASA4/PRKRIP1, KLF4, NRBF2, FGFR2, PIDD1, CCND1, RCCD1, ABHD8, and  
383 ZNF404<sup>7,9,10,13,17-22</sup>. However, for the large majority of the GWAS-identified breast cancer risk  
384 loci, the genes responsible for the associations remain unknown.

385

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

403

## 404 **Results**

### 405 **Gene expression prediction models**

406 The overall study design is shown in **Supplementary Figure 1**. We used transcriptome and  
407 high-density genotyping data from 67 women of European descent included in the Genotype-  
408 Tissue Expression (GTEx) project to build genetic models to predict RNA expression levels for

409 each of the genes expressed in normal breast tissues, by applying the elastic net method ( $\alpha=0.5$ )  
410 with ten-fold cross-validation. Genetically regulated expression was estimated for each gene  
411 using variants within a 2 MB window flanking the respective gene boundaries, inclusive. SNPs  
412 with a minor allele frequency of at least 0.05 and included in the HapMap Phase 2 subset were  
413 used for model building. Of the models built for 12,696 genes, 9,109 showed a prediction  
414 performance ( $R^2$ ) of at least 0.01 ( $\geq 10\%$  correlation between predicted and observed expression).  
415 For genes for which the expression could not be predicted well using this approach, we built  
416 models using only SNPs located in the promoter or enhancer regions, as predicted using three  
417 breast cell lines in the Roadmap Epigenomics Project/Encyclopedia of DNA Elements Project.  
418 This approach leverages information from functional genomics and reduces the number of  
419 variants for variable selection, and therefore potentially improving statistical power. This  
420 enabled us to build genetic models for additional 3,715 genes with  $R^2 \geq 0.01$ . **Supplementary**  
421 **Table 1** provides detailed information regarding the performance threshold and types of models  
422 built in this study. Overall, genes that were predicted with  $R^2 \geq 0.01$  in GTEx data were also  
423 predicted well in The Cancer Genome Atlas (TCGA) tumor-adjacent normal tissue data  
424 (correlation coefficient of 0.55 for  $R^2$  in two datasets; **Supplementary Figure 2**). Based on  
425 model performance in GTEx and TCGA, we prioritized 8,597 genes for analyses of the  
426 associations between predicted gene expression and breast cancer risk using the following  
427 criteria: 1) genes with a model prediction  $R^2$  of at least 0.01 in the GTEx set (10% correlation)  
428 and a Spearman's correlation coefficient of  $\geq 0.1$  in the external validation experiment using  
429 TCGA data, 2) genes with a prediction  $R^2$  of at least 0.09 (30% correlation) in the GTEx set  
430 regardless of their performance in the TCGA set, 3) genes with a prediction  $R^2$  of at least 0.01 in

431 the GTEx set (10% correlation) that could not be evaluated in the TCGA set because of a lack of  
432 data.

433

#### 434 **Association analyses of predicted gene expression with breast cancer risk**

435 Using the MetaXcan method<sup>26</sup>, we performed association analyses to evaluate predicted gene  
436 expression and breast cancer risk using the meta-analysis summary statistics of individual

437 genetic variants generated for 122,977 breast cancer cases and 105,974 controls of European

438 ancestry included in BCAC. For the majority of the tested genes, most of the SNPs selected for

439 prediction models were used for the association analyses (e.g.,  $\geq 95\%$  predicting SNPs used for

440  $83.8\%$  of the tested genes, and  $\geq 80\%$  predicting SNPs used for  $95.6\%$  of the tested genes).

441 Lambda 1,000 ( $\lambda_{1,000}$ ), a standardized estimate of the genomic inflation scaling to a study of

442 1,000 cases and 1,000 controls, was 1.004 in our study (Quantile-quantile (QQ) plot presented in

443 **Supplementary Figure 3 (A)**). Of the 8,597 genes evaluated in this study, we identified 179

444 genes whose predicted expression was associated with breast cancer risk at  $P < 1.05 \times 10^{-3}$ , a FDR-

445 corrected significance level (**Figure 1, Supplementary Table 2**). Of these, 48 showed a

446 significant association at the Bonferroni-corrected threshold of  $P \leq 5.82 \times 10^{-6}$  (**Figure 1, Tables 1-**

447 **3**), including 14 genes located at 11 loci that are 500 kb away from any of the risk variants

448 identified in previous GWAS of breast cancer risk (**Table 1**). An association between lower

449 predicted expression and increased breast cancer risk was detected for LRRC3B (3p24.1),

450 SPATA18 (4q12), UBD (6p22.1), MIR31HG (9p21.3), RIC8A (11p15.5), B3GNT1 (11q13.2),

451 GALNT16 (14q24.1) and MAN2C1 and CTD-2323K18.1 (15q24.2). Conversely, an association

452 between higher predicted expression and increased breast cancer risk was identified for ZSWIM5

453 (1p34.1), KLHDC10 (7q32.2), RP11-867G23.10 (11q13.2), RP11-218M22.1 (12p13.33) and

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

468

469 To determine whether the associations between predicted gene expression and breast cancer risk  
470 were independent of the association signals identified in previous GWAS, we performed  
471 conditional analyses adjusting for the GWAS-identified risk SNPs closest to the TWAS-  
472 identified gene (**Supplementary Table 4**)<sup>36</sup>. We found that the associations for 11 genes  
473 (LRRC3B, SPATA18, KLHDC10, MIR31HG, RIC8A, B3GNT1, RP11-218M22.1, MAN2C1,  
474 CTD-2323K18.1 (**Table 1**), ALK, CTD-3051D23.1 (**Table 2**)) remained statistically significant  
475 at  $P < 5.82 \times 10^{-6}$  (**Tables 1-3**). This suggests the expression of these genes may be associated with  
476 breast cancer risk independent of the GWAS-identified risk variant(s). For nine of the genes



477 (SPATA18, KLHDC10, MIR31HG, RIC8A, RP11-218M22.1, MAN2C1, CTD-2323K18.1 (**Table**  
478 **1**), ALK, and CTD-3051D23.1 (**Table 2**)), the significance level of the association remained  
479 essentially unchanged, suggesting these associations may be entirely independent of GWAS-  
480 identified association signals.

481  
482 Of the 131 genes showing a significant association at P values between  $5.82 \times 10^{-6}$  and  $1.05 \times 10^{-3}$   
483 (significant after FDR-correction but not Bonferroni-correction), 38 are located at GWAS-  
484 identified breast cancer risk loci ( $\pm 500$  kb of the index SNPs) (**Table 4**). Except for RP11-  
485 400F19.8, there was no evidence of heterogeneity in TWAS association ( $I^2 < 0.2$ ) across the  
486 iCOGS, OncoArray, and GWAS studies (**Supplementary Table 3**). After adjusting for the index  
487 SNPs, breast cancer associations for MTHFD1L, PVT1, RP11-123K19.1, FES, RP11-400F19.8,  
488 CTD-2538G9.5, and CTD-3216D2.5 remained significant at  $p \leq 1.05 \times 10^{-3}$ , again suggesting that  
489 the association of these genes with breast cancer risk may be independent of the GWAS-  
490 identified association signals (**Table 4**).

491  
492 For 41 of the 48 associated genes that reached the Bonferroni-corrected significant level, we  
493 obtained individual-level data from subjects included in the iCOGS (n=84,740) and OncoArray  
494 (n=112,133) datasets, which was 86% of the subjects included in the analysis using summary  
495 statistics (**Supplementary Table 5**). The results from the analysis using individual-level data  
496 were very similar to those described above using MetaXcan analyses (Pearson correlation of z-  
497 scores was 0.991 for iCOGS data and 0.994 for OncoArray data), although not all associations  
498 reached the Bonferroni-corrected significant level, possibly due to a smaller sample size  
499 (**Supplementary Table 5**). Conditional analyses using individual level data also revealed

500 consistent results compared with analyses using summary data. We found that for several genes  
501 within the same genomic region, their predicted expression levels were correlated with each  
502 other (**Tables 1-3**). The associations between predicted expression of PLEKHD1 and ZSWIM5  
503 and breast cancer risk were largely influenced by their corresponding closest risk variants  
504 identified in GWAS, although these risk variants are >500 kb away from these genes (**Table 1**).  
505 There were significant correlation of rs999737 and rs1707302 with genetically predicted  
506 expression of PLEKHD1 ( $r = -0.47$  in the OncoArray dataset and  $-0.48$  in the iCOGS dataset)  
507 and ZSWIM5 ( $r = 0.50$  in the OncoArray dataset and  $0.51$  in the iCOGS dataset), respectively.

508

#### 509 **INQUISIT algorithm scores for the identified genes**

510 For the 48 associated genes after Bonferroni correction, we assessed their integrated expression  
511 quantitative trait and in silico prediction of GWAS target (INQUISIT) scores<sup>7</sup> to assess whether  
512 there are other lines of evidence beyond the scope of eQTL for supporting our TWAS-identified  
513 genes as candidate target genes at GWAS-identified loci. The detailed methodology for  
514 INQUISIT scores have been described elsewhere<sup>7</sup>. In brief, a score for each gene-SNP pair is  
515 calculated across categories representing potential regulatory mechanisms - distal or proximal  
516 gene regulation (promoter). Features contributing to the score are based on functionally  
517 important genomic annotations such as chromatin interactions, transcription factor binding, and  
518 eQTLs. Compared with evidence from eQTL only, INQUISIT scores incorporate additional lines  
519 of evidence, including distal regulations. The INQUISIT scores for our identified genes are  
520 shown in **Supplementary Table 6**. Except for UBD with a very low score in the distal regulation  
521 category (0.05), none of the genes at novel loci (**Table 1**) showed evidence to be potential target  
522 genes for any of the GWAS-identified breast cancer susceptibility loci. This is interesting and

523 within the expectation since these genes may represent novel association signals. There was  
524 evidence suggesting that RP11-439A17.7, NUDT17, ANKRD34A, BTN3A2, AP006621.6,  
525 RPLP2, LRRC37A2, LRRC37A, KANSL1-AS1, CRHR1 and HAPLN4 listed in Table 2, and all  
526 eleven genes listed in Table 3, may be target genes for risk variants identified in GWAS at these  
527 loci (**Supplementary Table 6**). For NUDT17, ANKRD34A, RPLP2, LRRC37A2, LRRC37A,  
528 KANSL1-AS1, CRHR1, HAPLN4, KLHDC7A, ALS2CR12, CASP8, ATG10, ATP6AP1L,  
529 L3MBTL3, RMND1, SNX32, RCCD1, STXBP4 and ZNF404, the INQUISIT scores were not  
530 derived only from eQTL data, providing orthogonal support for these loci. For these loci, the  
531 associations of candidate causal SNPs with breast cancer risk may be mediated through these  
532 genes. This is in general consistent with the findings from the conditional analyses described  
533 above.

534

### 535 **Pathway enrichment analyses**

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

545

**546 Knockdown of predicted risk-associated genes in breast cells**

547 To assess the function of genes whose high levels of predicted expression were associated with  
548 increased breast cancer risk, we selected 13 genes for knockdown experiments in breast cells:  
549 ZSWIM5, KLHDC10, RP11-218M22.1 and PLEKHD1 (**Table 1**), UBLCP1, AP006621.6, RP11-  
550 467J12.4, CTD-3032H12.1 and RP11-15A1.7 (**Table 2**), and ALS2CR12, RMND1, STXBP4 and  
551 ZNF404 (**Table 3**). As negative controls, we selected B2M, ARHGDIA and ZAP70 using the  
552 following criteria: 1) at least 2 MB from any known breast cancer risk locus; 2) not an essential  
553 gene in breast cancer<sup>38,39</sup>; and 3) not predicted to be a target gene in INQUISIT. In addition, as  
554 positive controls, we included in the experiments PIDD1 (**Table 4**)<sup>7</sup>, NRBF2<sup>20</sup> and ABHD8<sup>22</sup>,  
555 which have been functionally validated as the target genes at breast cancer risk loci. We  
556 performed quantitative PCR (qPCR) on a panel of three ‘normal’ mammary epithelial and 15  
557 breast cancer cell lines to analyze their expression level (**Supplementary Figure 4 and**  
558 **Supplementary Table 8**). All 19 genes were expressed in the normal mammary epithelial line  
559 184A1<sup>40</sup> and the luminal breast cancer cell lines, MCF7 and T47D, so we used these cell lines  
560 for the proliferation assay, and MCF7 for the colony formation assay<sup>41</sup>. We also evaluated  
561 SNX32, ALK and BTN3A2 by qPCR, but they were not expressed in T47D and MCF7 cells;  
562 therefore they were not evaluated further. It was difficult to design siRNAs against RP11-  
563 867G23.1 and RP11-53O19.1 because they both have multiple transcripts with limited, GC-rich  
564 regions in common. We did not include RPLP2 because it is already known to be an essential  
565 gene for breast cancer survival<sup>42</sup>. Knockdown of the 19 tested genes was achieved by small short  
566 interfering RNA (siRNA) (**Supplementary Table 9**) and the knockdown efficiency was  
567 calculated in 184A1, MCF7 and T47D for each siRNA pair. Robust knockdown of the gene of

568 interests (GOI) was validated by qPCR with the majority of the siRNAs (**Supplementary Figure**  
569 **5**).

570

571 To evaluate the survival and proliferation ability of cells following gene interruption, we used an  
572 IncuCyte to quantify cell proliferation in real time and quantified the corrected proliferation of  
573 cells with knocking down of GOI in comparison to that of cells with non-target control (NTC)  
574 siRNA). As expected, knockdown of the three negative control genes (B2M, ARHGDI1 and  
575 ZAP70) did not significantly change cell proliferation in any of the three cell lines (**Figure 2A,**  
576 **Supplementary Figure 6**). However, with the exception of *UBLCP1*, *RMND1* and *STXBP4*,  
577 knockdown of all other genes (11 TWAS-identified genes along with two known genes, *ABHD8*  
578 and *NRBF2*) resulted in significantly decreased cell proliferation in 184A1 normal breast cells,  
579 with *KLHDC10*, *PLEKHD1*, *RP11-218M22.1*, *AP006621.6*, *ZNF404*, *RP11-467J12.4*, *CTD-*  
580 *3032H12.1* and *STXBP4* showing a similar effect in one or both cancer cell lines. Down-  
581 regulation of three lncRNAs (*RP11-218M22.1*, *RP11-467J12.4* and *CTD-3032H12.1*) resulted in  
582 significant reduction in cell proliferation in all three cell lines. We also evaluated the effect of  
583 inhibition of these genes on colony forming ability in MCF7 cells. Knockdown of the three  
584 negative control genes did not significantly affect colony forming efficiency (CFE). By contrast,  
585 knockdown of *PIDD1*, *RP11-15A1.7*, *RP11-218M22.1*, *AP006621.6*, *ZNF404*, *RP11-467J12.4*  
586 and *CTD-3032H12.1* resulted in significantly decreased colony forming efficiency in MCF7 cells  
587 compared to the NTC (**Figure 2B, Supplementary Figure 7**).

588

## 589 **Discussion**

590 This is the largest study to systematically evaluate associations of genetically predicted gene

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

599

600 It is possible that TWAS-identified genes may be associated with breast cancer risk through their  
601 correlation with disease causal genes. To determine the potential functional significance of  
602 TWAS-identified genes and provide evidence for causal inference, we knocked down 13 genes  
603 for which high predicted levels of expression were associated with an increased breast cancer  
604 risk, in one normal and two breast cancer cell lines, and measured the effect on proliferation and  
605 colony forming efficiency. Although there was some variation between cell lines, knockdown of  
606 11 of the 13 genes showed an effect in at least one cell line, particularly on proliferation in  
607 184A1 normal breast cells; the effects were strongest and most consistent for the lncRNAs,  
608 RP11-218M22.1, RP11-467J12.4 and CTD-3032H12.1. The observation of a more consistent  
609 effect in the normal breast cell line compared with the cancer cell lines is not surprising as cancer  
610 cell lines have increased capacity to handle gene interference through mutations which enhance  
611 cell survival. Rewiring of pathways and compensatory mechanisms is a hallmark of cancer.  
612 Knockdown of PIDD1, NRBF2 and ABHD8, for which breast cancer risk associated haplotypes  
613 have been shown to be associated with increased expression in reporter assays<sup>7,20,22</sup>, affected

614 either proliferation or colony forming efficiency, supporting the results from this study.  
615 Knockdown of UBLCP1 and RMND1 did not affect proliferation or colony formation but they  
616 could mediate breast cancer risk through other mechanisms.

617

618 Some of the genes with strong functional evidence from our study have been reported to have  
619 important roles in carcinogenesis. For example, RP11-467J12.4 (PR-lncRNA-1) is a p53-  
620 regulated lncRNA that modulates gene expression in response to DNA damage downstream of  
621 p53<sup>43</sup>. STXBP4 encodes Syntaxin binding protein 4, a scaffold protein that can stabilise and  
622 prevent degradation of an isoform of p63, a member of the p53 tumor suppressor family<sup>44</sup>.  
623 KLHDC10 encodes a member of the Kelch superfamily that can activate apoptosis signal-  
624 regulating kinase 1, contributing to oxidative stress-induced cell death<sup>45</sup>. Notably, another  
625 member of this superfamily, KLHDC7A, has recently been identified as the target gene at the  
626 1p36 breast cancer risk locus<sup>7</sup>.

627

628 SNX32, ALK and BTN3A2 are also likely susceptibility genes for breast cancer risk. However,  
629 their low or absent expression in our chosen breast cell lines prevented further functional  
630 analysis. SNX32 (Sorting Nexin 32) is not well characterized, but ALK (Anaplastic lymphoma  
631 kinase) copy number gain and overexpression have been reported in aggressive and metastatic  
632 breast cancers<sup>46</sup>. Therapeutic targeting of ALK rearrangement has significantly improved  
633 survival in advanced ALK-positive lung cancer<sup>47</sup>, making it an attractive target for breast and  
634 other cancers. BTN3A2 is a member of the B7/butyrophilin-like group of Ig superfamily  
635 receptors modulating the function of T-lymphocytes. While the exact role of BTN3A2 remains

636 unknown, over-expression of this gene in epithelial ovarian cancer is associated with higher  
637 infiltrating immune cells and a better prognosis<sup>48</sup>.

638

639 Our analyses identified multiple genes with reduced expression levels associated with increased  
640 breast cancer risk. Among them, LRRC3B and CASP8 are putative tumor suppressors in multiple  
641 cancers, including breast cancer. Leucine-rich repeat-containing 3B (LRRC3B) is a putative  
642 LRR-containing transmembrane protein, which is frequently inactivated via promoter  
643 hypermethylation leading to inhibition of cancer cell growth, proliferation, and invasion<sup>49</sup>.

644 CASP8 encodes a member of the cysteine-aspartic acid protease family, which play a central role  
645 in cell apoptosis. Previous studies have suggested that caspase-8 may act as a tumor suppressor  
646 in certain types of lung cancer and neuroblastoma, although this function has not yet been  
647 demonstrated in breast cancer. Notably, several large association studies have identified SNPs at  
648 the 2q33/CASP8 locus associated with increased breast cancer risk<sup>31,50</sup>. Consistent with our data,  
649 eQTL analyses showed that the risk alleles for breast cancer were associated with reduced  
650 CASP8 mRNA levels in both peripheral blood lymphocytes and normal breast tissue<sup>31</sup>.

651

652 For seven of the genes listed in Tables 1 and 2, we found some evidence from studies using  
653 tumor tissues, in vitro or in vivo experiments linking them to cancer risk (**Supplementary Table**  
654 **10**), although their association with breast cancer has not been previously demonstrated in human  
655 studies. For five of them, including LRRC3B, SPATA18, RIC8A, ALK and CRHR1, previous in  
656 vitro and in vivo experiments and human tissue studies showed a consistent direction of the  
657 association as demonstrated in our studies. For two other genes (UBD and MIR31HG), however,  
658 results from previous studies were inconsistent, reporting both potential promoting and inhibiting



659 effects on breast cancer development. Future studies are needed to evaluate functions of these  
660 genes.

661  
662 We included a large number of cases and controls in this study, providing strong statistical power  
663 for the association analysis. This large sample size enabled us to identify a large number of  
664 candidate breast cancer susceptibility genes, much larger than the number identified in a TWAS  
665 study with a sample size of about 20% of ours<sup>30</sup>. The previous study included subjects of  
666 different races, which could affect the results as linkage disequilibrium (LD) patterns differ by  
667 races. Of the five genes reported in that smaller TWAS that showed a suggestive association with  
668 breast cancer risk, the association for the RCCD1 gene was replicated in our study (**Table 3**).

669 The other four genes (ANKLE1, DHODH, ACAP1 and LRRC25) were not evaluated in our study  
670 because of unsatisfactory performance of our breast specific models for these genes which were  
671 built using the GTEx reference dataset including only female European descendants. In our  
672 study, the expression prediction model for ANKLE1 has a marginal performance in predicting  
673 gene expression ( $R^2=0.013$  in the GTEx). The model, however, did not perform well in the  
674 TCGA data. For ACAP1 and LRRC25, previous results for suggestive associations were based on  
675 blood tissue models.

676  
677 A substantial proportion of SNPs included in the OncoArray and iCOGS were selected from  
678 breast cancer GWAS and fine-mapping analyses, and thus these arrays were enriched for  
679 association signals with breast cancer risk. As a result, the overall  $\lambda$  value for the BCAC  
680 association analyses of individual variants is 1.26 after adjusting for population stratifications  
681 (QQ plot in **Supplementary Figure 3 (B)**)<sup>7</sup>. The  $\lambda$  value for the associations of the ~257,000

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

694

695 **The statistical power of our study is very large to detect associations for genes with a relatively**  
696 **high cis-heritability ( $h^2$ ) (**Supplementary Figure 8**). For example, our study has 80% statistical**  
697 **power to detect an association with breast cancer risk at  $P < 5.82 \times 10^{-6}$  with an OR of 1.07 or**  
698 **higher per one standard deviation increase (or decrease) in the expression level of genes with an**  
699  **$h^2$  of 0.1 or higher.** One limitation of our study is the small sample size for building gene  
700 expression prediction models, which may have affected the precision of model parameter  
701 estimates. The prediction performance ( $R^2$ ) for several of the genes identified in our study was  
702 not optimal, and thus additional research is needed to confirm our findings. We expect that  
703 models built with a larger sample size (and thus with more stable estimates of model parameters)  
704 will identify additional association signals. We used samples from women of European origin in

705 model building, given differences in gene expression patterns between males and females and in  
706 genetic architecture across ethnicities<sup>54</sup>. We also used gene expression data of tumor-adjacent  
707 normal tissue samples from European descendants in TCGA as an external validation step to  
708 prioritize genes for association analyses. Given potential somatic alterations in tumor-adjacent  
709 normal tissues, we retained all models showing a prediction performance ( $R^2$ ) of at least 0.09 in  
710 GTEx, regardless of their performance in TCGA. Not all genes have a significant hereditary  
711 component in expression regulation, and thus these genes could not be investigated in our study.  
712 For example, previous studies have provided strong evidence to support a significant role of the  
713 TERT, ESR1, CCND1, IGFBP5, TET2 and MRPS30 genes in the etiology of breast cancer.  
714 However, expression of these genes cannot be predicted well using the data from female  
715 European descendants included in the GTEx and thus they were not included in our association  
716 analyses. **Supplementary Table 11** summarizes the performance of prediction models and  
717 association results for breast cancer target genes reported previously at GWAS-identified loci.  
718  
719 In summary, our study has identified multiple gene candidates that can be further functionally  
720 characterized. By evaluating the associations of predicted gene expression levels with breast  
721 cancer risk, we provided evidence for the direction of the association for the identified genes.  
722 The silencing experiments we performed suggest that many of the genes identified by TWAS are  
723 likely to mediate risk of breast cancer by affecting proliferation or colony forming efficiency,  
724 two of the hallmarks of cancer. Further investigation of genes identified in our study will provide  
725 additional insight into the biology and genetics of breast cancer.

726

727 **Methods**

**728 Building of gene expression prediction models**

729 We used transcriptome and high-density genotyping data from the Genotype-Tissue Expression  
730 (GTEx) study to establish prediction models for genes expressed in normal breast tissues. Details  
731 of the GTEx have been described elsewhere<sup>55</sup>. Genomic DNA samples obtained from study  
732 subjects included in the GTEx were genotyped using Illumina OMNI 5M or 2.5M SNP Array  
733 and RNA samples from 51 tissue sites were sequenced to generate transcriptome profiling data.  
734 Genotype data were processed according to the GTEx protocol  
735 (<http://www.gtexportal.org/home/documentationPage>). SNPs with a call rate < 98%, with  
736 differential missingness between the two array experiments (5M/2.5M Arrays), with Hardy-  
737 Weinberg equilibrium p-value < 10<sup>-6</sup> (among subjects of European ancestry), or showing batch  
738 effects were excluded. One Klinefelter individual, three related individuals, and a chromosome  
739 17 trisomy individual were also excluded. The genotype data were imputed to the Haplotype  
740 Reference Consortium reference panel<sup>56</sup> using Minimac3 for imputation and SHAPEIT for  
741 prephasing<sup>57,58</sup>. SNPs with high imputation quality ( $r^2 \geq 0.8$ ), minor allele frequency (MAF)  $\geq$   
742 0.05, and included in the HapMap Phase 2 version, were used to build expression prediction  
743 models. For gene expression data, we used Reads Per Kilobase per Million (RPKM) units from  
744 RNA-SeQC<sup>59</sup>. Genes with a median expression level of 0 RPKM across samples were removed,  
745 and the RPKM values of each gene were log2 transformed. We performed quantile normalization  
746 to bring the expression profile of each sample to the same scale, and performed inverse quantile  
747 normalization for each gene to map each set of expression values to a standard normal. We  
748 adjusted for the top ten principal components (PCs) derived from genotype data and the top 15  
749 probabilistic estimation of expression residuals (PEER) factors to correct for batch effects and  
750 experimental confounders in model building<sup>60</sup>. Genetic and transcriptome data from 67 female

751 subjects of European descent without a prior breast cancer diagnosis were used to build gene  
752 expression prediction models for this study.

753

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

765

766 For genes that cannot be predicted well using the above approach, we built models using only  
767 SNPs located in predicted promoter or enhancer regions in breast cell lines. This approach  
768 reduces the number of variants for model building, and thus potentially improves model  
769 accuracy, by increasing the ratio of sample size to effective degrees of freedom.

770 SNP-level annotation data in three breast cell lines, namely, Breast Myoepithelial Primary Cells  
771 (E027), Breast variant Human Mammary Epithelial Cells (vHMEC) (E028), and HMEC  
772 Mammary Epithelial Primary Cells (E119) in the Roadmap Epigenomics Project/Encyclopedia  
773 of DNA Elements Project<sup>16</sup>, were downloaded from

774 <http://archive.broadinstitute.org/mammals/haploreg/data/> (Version 4.0, assessed on December 6,  
775 2016). SNPs in regions classified as promoters (TssA, TssAFlnk), enhancers (Enh, EnhG), or  
776 regions with both promoter and enhancer signatures (ExFlnk) according to the core 15 chromatin  
777 state model<sup>16</sup> in at least one of the cell lines were retained as input SNPs for model building.

778

### 779 **Evaluating performance of gene expression prediction models using The Cancer Genome** 780 **Atlas (TCGA) data**

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

788

### 789 **Evaluating statistical power for association tests**

790 We conducted a simulation analysis to assess the power of our TWAS analysis. Specifically, we  
791 set the number of cases and controls to be 122,977 and 105,974, respectively, and generated the  
792 gene expression levels from the empirical distribution of predicted gene expression levels in the  
793 BCAC. We calculated statistical power at  $P < 5.82 \times 10^{-6}$  (the significance level used in our  
794 TWAS) according to cis-heritability ( $h^2$ ) which we aim to capture using gene expression  
795 prediction models ( $R^2$ ). The results based on 1000 replicates are summarized in **Supplementary**  
796 **Figure 8**. Based on the power calculation, our TWAS analysis has 80% power to detect a

797 minimum odds ratio of 1.11, 1.07, 1.05, 1.04, or 1.03 for breast cancer risk per one standard  
 798 deviation increase (or decrease) in the expression level of a gene whose cis-heritability is 5%,  
 799 10%, 20%, 40%, or 60%, respectively.

800

### 801 **Association analyses of predicted gene expression with breast cancer risk**

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

810

811 To identify novel breast cancer susceptibility loci and genes, the MetaXcan method, as described  
 812 elsewhere, was used for the association analyses<sup>26</sup>. Briefly, the formula:

$$813 \quad Z_g \approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)}$$

814 was used to estimate the Z-score of the association between predicted expression and breast  
 815 cancer risk. Here  $w_{lg}$  is the weight of SNP  $l$  for predicting the expression of gene  $g$ ,  $\hat{\beta}_l$  and  
 816  $\text{se}(\hat{\beta}_l)$  are the GWAS association regression coefficient and its standard error for SNP  $l$ , and  $\hat{\sigma}_l$   
 817 and  $\hat{\sigma}_g$  are the estimated variances of SNP  $l$  and the predicted expression of gene  $g$  respectively.  
 818 Therefore, the weights for predicting gene expression, GWAS summary statistics results, and

819 correlations between model predicting SNPs are the input variables for the MetaXcan analyses.  
820 For this study we estimated correlations between SNPs included in the prediction models using  
821 the phase 3, 1000 Genomes Project data focusing on European population.

822

823 For the association analysis, we used the summary statistics data of genetic variants associated  
824 with breast cancer risk generated in 122,977 breast cancer patients and 105,974 controls of  
825 European ancestry from the Breast Cancer Association Consortium (BCAC). The details of the  
826 BCAC have been described elsewhere<sup>7,9,13,64,65</sup>. Briefly, 46,785 breast cancer cases and 42,892  
827 controls of European ancestry were genotyped using a custom Illumina iSelect genotyping array  
828 (iCOGS) containing ~211,155 variants. A further 61,282 cases and 45,494 controls of European  
829 ancestry were genotyped using the OncoArray including 570,000 SNPs

830 (<http://epi.grants.cancer.gov/oncoarray/>). Also included in this analysis were data from nine  
831 GWAS studies including 14,910 breast cancer cases and 17,588 controls of European ancestry.

832 Genotype data from iCOGS, OncoArray and GWAS were imputed using the October 2014  
833 release of the 1000 Genomes Project data as reference. Genetic association results for breast  
834 cancer risk were combined using inverse variance fixed effect meta-analyses<sup>7</sup>. For our study,  
835 only SNPs with imputation  $r^2 \geq 0.3$  were used. All participating BCAC studies were approved by  
836 their appropriate ethics review boards. This study was approved by the BCAC Data Access  
837 Coordination Committee.

838

839 Lambda 1,000 ( $\lambda_{1,000}$ ) was calculated to represent a standardized estimate of the genomic

840 inflation scaling to a study of 1,000 cases and 1,000 controls, using the following formula:

841  $\lambda_{1,000} = 1 + (\lambda_{\text{obs}} - 1) \times (1/n_{\text{cases}} + 1/n_{\text{controls}}) / (1/1,000_{\text{cases}} + 1/1,000_{\text{controls}})$ <sup>66,67</sup>. We used a Bonferroni



842 corrected p threshold of  $5.82 \times 10^{-6}$  ( $0.05/8,597$ ) to determine a statistically significant association  
843 for the primary analyses. To identify additional gene candidates at previously identified  
844 susceptibility loci, we also used a false discovery rate (FDR) corrected p threshold of  $1.05 \times 10^{-3}$   
845 ( $\text{FDR} \leq 0.05$ ) to determine a significant association. Associated genes with an expression of  $>0.1$   
846 RPKM in less than 10 individuals in GTEx data were excluded as the corresponding prediction  
847 models may not be stable.

848

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

856

857 For 41 identified associated genes at the Bonferroni-corrected threshold, we also performed  
858 analyses using individual level data in iCOGS ( $n=84,740$ ) and OncoArray ( $n=112,133$ ) datasets.  
859 We generated predicted gene expression using predicting SNPs, and then assessed the  
860 association between predicted gene expression and breast cancer risk adjusting for study and  
861 nine principal components in iCOGS dataset, and country and the first ten principal components  
862 in OncoArray dataset. Conditional analyses adjusting for index SNPs were performed to assess  
863 potential influence of reported index SNPs on the association between predicted gene expression  
864 and breast cancer risk. Furthermore, we evaluated whether the predicted expression levels of

865 genes within a same genomic region were correlated with each other by using the OncoArray  
866 data.

867

### 868 **INQUISIT algorithm scores for TWAS-identified genes**

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

888

**889 Functional enrichment analysis using Ingenuity Pathway Analysis (IPA)**

890 We performed functional enrichment analysis for the identified protein-coding genes reaching  
891 Bonferroni corrected association threshold. To assess potential functionality of the identified  
892 lncRNAs, we examined their co-expressed protein-coding genes determined using expression  
893 data of normal breast tissue of European females in GTEx. Spearman's correlations between  
894 protein-coding genes and identified lncRNAs of  $\geq 0.4$  or  $\leq -0.4$  were used to indicate a high co-  
895 expression. Canonical pathways, top associated diseases and biofunctions, and top networks  
896 associated with genes of interest were estimated using IPA software<sup>37</sup>.

897

**898 Gene expression in breast cell lines**

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

909

**910 Short interfering RNA (siRNA) silencing**

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

915

### 916 **Proliferation and colony formation assays**

917 For proliferation assays, MCF7 and T47D cells were trypsinized at 16 hours post-transfection  
918 and seeded into 24 well plates to achieve ~10% confluency. Phase-contrast images were  
919 collected with IncuCyte ZOOM (Essen Bioscience) for seven days. Duplicate samples were  
920 assessed for each GOI siRNA transfected cells along with non-target control si (NTCsi) treated  
921 cells in the same plate. 184A1 cells were reverse-transfected in 96 well plates to achieve 50%  
922 confluence at 8 hours after transfection. Two independent experiments were carried out for all  
923 siRNAs in all three cell lines. Each cell proliferation time-course was normalized to the baseline  
924 confluency and analyzed in GraphPad Prism. The area under the curve was calculated for each  
925 concentration (n=4) and used to calculate corrected proliferation (Corrected proliferation % =  
926  $100 \pm (\text{relative proliferation in indicated siRNA} - \text{proliferation in NTC siRNA}) / \text{knockdown}$   
927 efficiency (“+” if the GOI promotes proliferation and “-” if it inhibits proliferation)). For each  
928 gene, results from two siRNAs in two independent experiments were averaged and summarized  
929 in **Figure 2** and **Supplementary Figure 6**. For colony formation assays; the same number of  
930 GOI siRNA transfected MCF7 cells was seeded in 6 well plates at 16 hours after transfection to  
931 assay colony forming efficiency at two weeks. All siRNA-treated cells were seeded in duplicate.  
932 Colonies (defined to consist of at least 50 cells) were fixed with methanol, stained with crystal  
933 violet (0.5% w/v), scanned and counted using ImageJ as batch analysis by a self-defined plug-in

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

938

### 939 **Data availability**

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

947

### 948 **Code availability**

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

950

### 951 **Acknowledgements**

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

957 Hae Kyung Im of University of Chicago for her help. The data analyses were conducted using  
958 the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University.  
959 This project at Vanderbilt University Medical Center was supported in part by grants  
960 R01CA158473 and R01CA148677 from the U.S. National Institutes of Health as well as funds  
961 from Anne Potter Wilson endowment. Dr. Wu was supported by the Vanderbilt Molecular and  
962 Genetic Epidemiology of Cancer (MAGEC) training program funded by the US National Cancer  
963 Institute grant R25 CA160056 (PI: X.-O. Shu). Genotyping of the OncoArray was principally  
964 funded from three sources: the PERSPECTIVE project, funded from the Government of Canada  
965 through Genome Canada and the Canadian Institutes of Health Research, the Ministère de  
966 *l'Économie, de la Science et de l'Innovation du Québec* through Genome Québec, and the  
967 Quebec Breast Cancer Foundation; the NCI Genetic Associations and Mechanisms in Oncology  
968 (GAME-ON) initiative and Discovery, Biology and Risk of Inherited Variants in Breast Cancer  
969 (DRIVE) project (NIH Grants U19 CA148065 and X01HG007492); and Cancer Research UK  
970 (C1287/A10118 and C1287/A16563). **BCAC** is funded by Cancer Research UK  
971 [C1287/A16563], by the European Community's Seventh Framework Programme under grant  
972 agreement 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union's Horizon  
973 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and  
974 634935 (BRIDGES). Genotyping of the iCOGS array was funded by the European Union  
975 (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of  
976 Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program, and the  
977 Ministry of Economic Development, Innovation and Export Trade of Quebec – grant # PSR-  
978 SIIRI-701. Combining the GWAS data was supported in part by The National Institute of Health  
979 (NIH) Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE, part of the GAME-

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

982

### 983 **Author Contributions**

984 W.Z. and J.L. conceived the study. L.W. contributed to the study design, and performed  
985 statistical analyses. L.W., W.Z. and G.C.-T. wrote the manuscript with significant contributions  
986 from W.S., J.L., X.G., and S.L.E.. W.S. performed the *in vitro* experiments. G.C.-T. directed the  
987 *in vitro* experiments. X.G. contributed to the model building and pathway analyses. J.B.  
988 contributed to the bioinformatics analyses. F.A.-E., E.R., and S.L.E. contributed to the *in vitro*  
989 experiments. Y. L. and C. Z. contributed to the model building. K.M., M.K.B., X.-O.S., Q.W.,  
990 J.D., B.L., C.Z., H.F., A.G., R.T.B., A.M.D., P.D.P.P., J.S., R.L.M., P.K., and D.F.E, contributed  
991 to manuscript revision, statistical analyses and/or BCAC data management. I.L.A., H.A.-C.,  
992 V.A., K.J.A., P.L.A., M. Barrdahl, C.B., M.W.B., J.B., M. Bermisheva, C.B., N.V.B., S.E.B., H.  
993 Brauch, H. Brenner, L.B., P.B., S.Y.B., B.B., Q.C., T.C., F.C., B.D.C., J.E.C., J.C.-C., X.C., T.-  
994 Y.D.C., H.C., C.L.C., NBCS Collaborators, M.C., S.C., F.J.C., D.C., A.C., S.S.C., J.M.C., K.C.,  
995 M.B.D., P.D., K.F.D., T.D., I.d.S.S., M. Dumont, M. Dwek, D.M.E., U.E., H.E., C.E., M.E.,  
996 L.F., P.A.F., J.F., D.F.-J., O.F., H.F., L.F., M. Gabrielson, M.G.-D., S.M.G., M.G.-C., M.M.G.,  
997 M. Ghossaini, G.G.G., M.S.G., D.E.G., A.G.-N., P.G., E. Hahnen, C.A.H., N.H., P. Hall, E.  
998 Hallberg, U.H., P. Harrington, A. Hein, B.H., P. Hillemanns, A. Hollestelle, R.N.H., J.L.H.,  
999 G.H., K.H., D.J.H., A.J., W.J., E.M.J., N.J., K.J., M.E.J., A. Jung, R.K., M.J.K., E.K., V.-M.K.,  
1000 V.N.K., D.L., L.L.M., J. Li, S.L., J. Lissowska, W.-Y.L., S.Loibl, J.L., C.L., M.P.L., R.J.M.,  
1001 T.M., I.M.K., A. Mannermaa, J.E.M., S.M., D.M., H.M.-H., A. Meindl, U.M., J.M., A.M.M.,  
1002 S.L.N., H.N., P.N., S.F.N., B.G.N., O.I.O., J.E.O., H.O., P.P., J.P., D.P.-K., R.P., N.P., K.P.,

1003 B.R., P.R., N.R., G.R., H.S.R., V.R., A. Romero, J.R., A. Rudolph, E.S., D.P.S, E.J.S., M.K.S.,  
1004 R.K.S., A.S., R.J.S., C. Scott, S.S., M.S., M.J.S., A.S., M.C.S., J.J.S., J.S., H.S., A.J.S., R.T.,  
1005 W.T., J.A.T., M.B.T., D.C.T., A.T., K.T., R.A.E.M.T., D.T., T.T., M.U., C.V., D.V.D.B., D.V.,  
1006 Q.W., C.R.W., C.W., A.S.W., H.W., W.C.W., R.W., A.W., L.X., X.R.Y., A.Z., E.Z.,  
1007 kConFab/AOCS Investigators contributed to the collection of the data and biological samples for  
1008 the original BCAC studies. All authors have reviewed and approved the final manuscript.

1009

1010 **Competing financial interests**

1011 The authors declare no competing financial interests.

1012

1013



1014 **References**

- 1015 1. Kamangar, F., Dores, G.M. & Anderson, W.F. Patterns of cancer incidence, mortality,  
1016 and prevalence across five continents: defining priorities to reduce cancer disparities in  
1017 different geographic regions of the world. *J Clin Oncol* **24**, 2137-50 (2006).
- 1018 2. Beggs, A.D. & Hodgson, S.V. Genomics and breast cancer: the different levels of  
1019 inherited susceptibility. *Eur J Hum Genet* **17**, 855-6 (2009).
- 1020 3. Southey, M.C. et al. PALB2, CHEK2 and ATM rare variants and cancer risk: data from  
1021 COGS. *J Med Genet* (2016).
- 1022 4. Nathanson, K.L., Wooster, R. & Weber, B.L. Breast cancer genetics: what we know and  
1023 what we need. *Nat Med* **7**, 552-6 (2001).
- 1024 5. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series  
1025 of breast cancer cases. Anglian Breast Cancer Study Group. *Br J Cancer* **83**, 1301-8  
1026 (2000).
- 1027 6. Milne, R.L. et al. Identification of ten variants associated with risk of estrogen-receptor-  
1028 negative breast cancer. *Nat Genet* **49**, 1767-1778 (2017).
- 1029 7. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci.  
1030 *Nature* **551**, 92-94 (2017).
- 1031 8. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with  
1032 breast cancer risk. *Nat Genet* **45**, 353-61, 361e1-2 (2013).
- 1033 9. Michailidou, K. et al. Genome-wide association analysis of more than 120,000  
1034 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373-80  
1035 (2015).
- 1036 10. Cai, Q. et al. Genome-wide association analysis in East Asians identifies breast cancer  
1037 susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat Genet* **46**, 886-90 (2014).
- 1038 11. Zheng, W. et al. Common genetic determinants of breast-cancer risk in East Asian  
1039 women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. *Hum*  
1040 *Mol Genet* **22**, 2539-50 (2013).
- 1041 12. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with  
1042 breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological  
1043 evidence. *Lancet Oncol* **12**, 477-88 (2011).
- 1044 13. French, J.D. et al. Functional variants at the 11q13 risk locus for breast cancer regulate  
1045 cyclin D1 expression through long-range enhancers. *Am J Hum Genet* **92**, 489-503  
1046 (2013).
- 1047 14. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide  
1048 association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7  
1049 (2009).
- 1050 15. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome.  
1051 *Nature* **489**, 57-74 (2012).
- 1052 16. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human  
1053 epigenomes. *Nature* **518**, 317-30 (2015).
- 1054 17. Dunning, A.M. et al. Breast cancer risk variants at 6q25 display different phenotype  
1055 associations and regulate ESR1, RMND1 and CCDC170. *Nat Genet* **48**, 374-86 (2016).
- 1056 18. Ghousaini, M. et al. Evidence that breast cancer risk at the 2q35 locus is mediated  
1057 through IGFBP5 regulation. *Nat Commun* **4**, 4999 (2014).

- 1058 19. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci.  
1059 Cell **152**, 633-41 (2013).
- 1060 20. Darabi, H. et al. Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer  
1061 Risk Locus Regulate NRBF2 Expression. Am J Hum Genet **97**, 22-34 (2015).
- 1062 21. Glubb, D.M. et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least  
1063 three independent risk variants regulating MAP3K1. Am J Hum Genet **96**, 5-20 (2015).
- 1064 22. Lawrenson, K. et al. Functional mechanisms underlying pleiotropic risk alleles at the  
1065 19p13.1 breast-ovarian cancer susceptibility locus. Nat Commun **7**, 12675 (2016).
- 1066 23. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence.  
1067 Nat Genet **47**, 955-61 (2015).
- 1068 24. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-  
1069 wide association summary statistics. Nat Genet **47**, 1228-35 (2015).
- 1070 25. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants  
1071 across 11 common diseases. Am J Hum Genet **95**, 535-52 (2014).
- 1072 26. Barbeira, A.N. et al. Exploring the phenotypic consequences of tissue specific gene  
1073 expression variation inferred from GWAS summary statistics. bioRxiv (2017).
- 1074 27. Gamazon, E.R. et al. A gene-based association method for mapping traits using reference  
1075 transcriptome data. Nat Genet **47**, 1091-8 (2015).
- 1076 28. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association  
1077 studies. Nat Genet **48**, 245-52 (2016).
- 1078 29. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts  
1079 complex trait gene targets. Nat Genet **48**, 481-7 (2016).
- 1080 30. Hoffman, J.D. et al. Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes  
1081 associated with breast cancer risk. PLoS Genet **13**, e1006690 (2017).
- 1082 31. Lin, W.Y. et al. Identification and characterization of novel associations in the  
1083 CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. Hum Mol Genet **24**,  
1084 285-98 (2015).
- 1085 32. Camp, N.J. et al. Discordant Haplotype Sequencing Identifies Functional Variants at the  
1086 2q33 Breast Cancer Risk Locus. Cancer Res **76**, 1916-25 (2016).
- 1087 33. Li, Q. et al. Expression QTL-based analyses reveal candidate causal genes and loci across  
1088 five tumor types. Hum Mol Genet **23**, 5294-302 (2014).
- 1089 34. Caswell, J.L. et al. Multiple breast cancer risk variants are associated with differential  
1090 transcript isoform expression in tumors. Hum Mol Genet **24**, 7421-31 (2015).
- 1091 35. Darabi, H. et al. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs,  
1092 genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Sci  
1093 Rep **6**, 32512 (2016).
- 1094 36. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics  
1095 identifies additional variants influencing complex traits. Nat Genet **44**, 369-75, S1-3  
1096 (2012).
- 1097 37. Kramer, A., Green, J., Pollard, J., Jr. & Tugendreich, S. Causal analysis approaches in  
1098 Ingenuity Pathway Analysis. Bioinformatics **30**, 523-30 (2014).
- 1099 38. Koh, J.L. et al. COLT-Cancer: functional genetic screening resource for essential genes  
1100 in human cancer cell lines. Nucleic Acids Res **40**, D957-63 (2012).
- 1101 39. Marcotte, R. et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells.  
1102 Cancer Discov **2**, 172-89 (2012).

- 1103 40. Walen, K.H. & Stampfer, M.R. Chromosome analyses of human mammary epithelial  
1104 cells at stages of chemical-induced transformation progression to immortality. *Cancer*  
1105 *Genet Cytogenet* **37**, 249-61 (1989).
- 1106 41. Treszezamsky, A.D. et al. BRCA1- and BRCA2-deficient cells are sensitive to etoposide-  
1107 induced DNA double-strand breaks via topoisomerase II. *Cancer Res* **67**, 7078-81 (2007).
- 1108 42. Marcotte, R. et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells.  
1109 *Cancer Discov* **2**, 172-189 (2012).
- 1110 43. Sanchez, Y. et al. Genome-wide analysis of the human p53 transcriptional network  
1111 unveils a lncRNA tumour suppressor signature. *Nat Commun* **5**, 5812 (2014).
- 1112 44. Li, Y., Peart, M.J. & Prives, C. Stxbp4 regulates DeltaNp63 stability by suppression of  
1113 RACK1-dependent degradation. *Mol Cell Biol* **29**, 3953-63 (2009).
- 1114 45. Sekine, Y. et al. The Kelch repeat protein KLHDC10 regulates oxidative stress-induced  
1115 ASK1 activation by suppressing PP5. *Mol Cell* **48**, 692-704 (2012).
- 1116 46. Kim, M.H. et al. Anaplastic lymphoma kinase gene copy number gain in inflammatory  
1117 breast cancer (IBC): prevalence, clinicopathologic features and prognostic implication.  
1118 *PLoS One* **10**, e0120320 (2015).
- 1119 47. Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer. *N Engl J Med*  
1120 **373**, 1582 (2015).
- 1121 48. Le Page, C. et al. BTN3A2 expression in epithelial ovarian cancer is associated with  
1122 higher tumor infiltrating T cells and a better prognosis. *PLoS One* **7**, e38541 (2012).
- 1123 49. Kan, L. et al. LRRC3B is downregulated in non-small-cell lung cancer and inhibits  
1124 cancer cell proliferation and invasion. *Tumour Biol* **37**, 1113-20 (2016).
- 1125 50. Cox, A. et al. A common coding variant in CASP8 is associated with breast cancer risk.  
1126 *Nat Genet* **39**, 352-8 (2007).
- 1127 51. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet*  
1128 **19**, 807-12 (2011).
- 1129 52. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height.  
1130 *Nature* **542**, 186-190 (2017).
- 1131 53. Turcot, V. et al. Protein-altering variants associated with body mass index implicate  
1132 pathways that control energy intake and expenditure in obesity. *Nat Genet* **50**, 26-41  
1133 (2018).
- 1134 54. Mele, M. et al. Human genomics. The human transcriptome across tissues and  
1135 individuals. *Science* **348**, 660-5 (2015).
- 1136 55. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot  
1137 analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
- 1138 56. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat*  
1139 *Genet* **48**, 1279-83 (2016).
- 1140 57. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for  
1141 thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
- 1142 58. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation  
1143 method for the next generation of genome-wide association studies. *PLoS Genet* **5**,  
1144 e1000529 (2009).
- 1145 59. DeLuca, D.S. et al. RNA-SeQC: RNA-seq metrics for quality control and process  
1146 optimization. *Bioinformatics* **28**, 1530-2 (2012).

- 1147 60. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of  
1148 expression residuals (PEER) to obtain increased power and interpretability of gene  
1149 expression analyses. *Nat Protoc* **7**, 500-7 (2012).
- 1150 61. Guo, X., Lin, M., Rockowitz, S., Lachman, H.M. & Zheng, D. Characterization of human  
1151 pseudogene-derived non-coding RNAs for functional potential. *PLoS One* **9**, e93972  
1152 (2014).
- 1153 62. Casbas-Hernandez, P. et al. Tumor intrinsic subtype is reflected in cancer-adjacent tissue.  
1154 *Cancer Epidemiol Biomarkers Prev* **24**, 406-14 (2015).
- 1155 63. Huang, X., Stern, D.F. & Zhao, H. Transcriptional Profiles from Paired Normal Samples  
1156 Offer Complementary Information on Cancer Patient Survival--Evidence from TCGA  
1157 Pan-Cancer Data. *Sci Rep* **6**, 20567 (2016).
- 1158 64. Ghousaini, M. et al. Genome-wide association analysis identifies three new breast  
1159 cancer susceptibility loci. *Nat Genet* **44**, 312-8 (2012).
- 1160 65. Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative-  
1161 specific breast cancer risk loci. *Nat Genet* **45**, 392-8, 398e1-2 (2013).
- 1162 66. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-  
1163 1004 (1999).
- 1164 67. Freedman, M.L. et al. Assessing the impact of population stratification on genetic  
1165 association studies. *Nat Genet* **36**, 388-93 (2004).
- 1166 68. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles  
1167 of chromatin looping. *Cell* **159**, 1665-80 (2014).
- 1168 69. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in  
1169 human cells. *Proc Natl Acad Sci U S A* **111**, E2191-9 (2014).
- 1170 70. Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage  
1171 disequilibrium dictate levels of gene expression to confer susceptibility to common traits.  
1172 *Genome Res* **24**, 1-13 (2014).
- 1173 71. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-  
1174 47 (2013).
- 1175 72. Consortium, F. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462-70  
1176 (2014).
- 1177
- 1178
- 1179

1180 **Figure Legends**1181 **Figure 1. Manhattan plot of association results from the breast cancer transcriptome-wide**1182 **association study.** The red line represents  $P = 5.82 \times 10^{-6}$ . The blue line represents  $P =$ 1183  $1.00 \times 10^{-3}$ .

1184

1185 **Figure 2. Heat maps of proliferation and colony formation efficiency in breast cells. (A)**

1186 184A1, MCF7 or T47D cells were transfected with indicated siRNAs over seven days and phase-

1187 contrast images collected using an IncuCyte ZOOM. Each cell proliferation time-course was

1188 normalized to the baseline confluency and analyzed using GraphPad Prism. Corrected

1189 proliferation % =  $100 \pm$  (relative proliferation in indicated siRNA - proliferation in control1190 siRNA (consi))/knockdown efficiency. **(B)** MCF7 cells were transfected with indicated siRNAs,

1191 then reseeded after 16 hours for colony formation (CF) assay. At day 14, colonies were fixed

1192 with methanol, stained with crystal violet, scanned and batch analyzed by ImageJ. Corrected CF

1193 efficiency (CFE) % =  $100 \pm$  (relative CFE in indicated siRNA - CFE in control siRNA

1194 (consi))/knockdown efficiency. Error bars, SD (N=2). P-values were determined by one-way

1195 ANOVA followed by Dunnett's multiple comparisons test: \*P-value &lt; 0.05. NTC: non-target

1196 control.

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1 **Table 1.** Fourteen expression-trait associations for genes located at genomic loci at least 500 kb away from any GWAS-identified  
 2 breast cancer risk variants  
 3

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

4  
 5 <sup>a</sup> Genes that were siRNA-silenced for functional assays are bolded; SNPs used to predict gene expression are listed in the Supplementary Table 13

6 <sup>b</sup> Protein: protein coding genes; lncRNA: long non-coding RNAs; transcript: processed transcript

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

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

11 <sup>e</sup> Use of COJO method<sup>36</sup>

12 <sup>f</sup> Predicted expression of MAN2C1 and CTD-2323K18.1 was correlated (spearman R=0.76)

13  
 14  
 15

1 **Table 2.** Twenty-three expression-trait associations for genes located at genomic loci within 500 kb of any previous GWAS-identified  
 2 breast cancer risk variants but not yet implicated as target genes of risk variants<sup>#</sup>  
 3

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

4 <sup>#</sup> not yet reported from eQTL and/or functional studies as target genes of GWAS-identified risk variants and not harbor GWAS or fine-mapping  
 5 identified risk variants

6 <sup>a</sup> Genes that were siRNA-silenced for functional assays are bolded; SNPs used to predict gene expression are listed in the Supplementary Table 13

- 1 <sup>b</sup> Protein: protein coding genes; lncRNA: long non-coding RNAs  
2 <sup>c</sup> P value: nominal P value from association analysis; the threshold after Bonferroni correction of 8,597 tests ( $0.05/8,597=5.82\times 10^{-6}$ ) was used; R<sup>2</sup>:  
3 prediction performance (R<sup>2</sup>) derived using GTEx data  
4 <sup>d</sup> Risk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and  
5 their distances to the genes are presented in the **Supplementary Table 4**  
6 <sup>e</sup> Use of COJO method<sup>36</sup>; all index SNPs in the corresponding region were adjusted in the conditional analyses  
7 <sup>f</sup> Predicted expression of RP11-73O6.3 and L3MBTL3 was correlated (spearman R=0.88)  
8 <sup>g</sup> Predicted expression of AP006621.6 and RPLP2 was correlated; predicted expression of LRRC37A, KANSL1-AS1, and CRHR1 was correlated  
9 (spearman R>0.1)  
10 <sup>h</sup> Predicted expression of RP11-15A1.7 and ZNF404 was correlated (spearman R=0.64)  
11



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

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

4  
 5 <sup>a</sup> Genes that were siRNA silenced for functional assays are bolded; SNPs used to predict gene expression are listed in the Supplementary Table 13

6 <sup>b</sup> Protein: protein coding genes; lncRNA: long non-coding RNAs; NA: not available

7 <sup>c</sup> P value: nominal P value from association analysis; the threshold after Bonferroni correction of 8,597 tests ( $0.05/8,597=5.82 \times 10^{-6}$ ) was used; R<sup>2</sup>:  
 8 prediction performance (R<sup>2</sup>) derived using GTEx data .

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

11 <sup>e</sup> Use of COJO method<sup>36</sup>; all index SNPs in the corresponding region were adjusted for the conditional analyses

12 <sup>f</sup> -: inverse association; +: positive association; mixed: both inverse and positive associations reported; NA: not available

13 <sup>g</sup> Predicted expression of L3MBTL3 and RP11-73O6.3 was correlated (spearman R=0.88)

14 <sup>h</sup> Predicted expression of ZNF404 and RP11-15A1.7 was correlated (spearman R=0.64)

15

16

17

18

1 **Table 4.** Genes at GWAS-identified breast cancer risk loci ( $\pm 500$ kb of the index SNPs) whose predicted expression levels were  
 2 associated with breast cancer risk at p-values between  $5.82 \times 10^{-6}$  and  $1.05 \times 10^{-3}$  (FDR corrected p-value  $\leq 0.05$ )  
 3

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

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

1

2 <sup>a</sup> Protein: protein coding genes; lncRNA: long non-coding RNAs; transcript: processed transcript3 <sup>b</sup>P value: nominal P value from association analysis; R<sup>2</sup>: prediction performance derived using GTEx data.4 <sup>c</sup> Risk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and  
5 their distances to the genes are presented in the **Supplementary Table 4**6 <sup>d</sup> Use of COJO method<sup>36</sup>; all index SNPs in the corresponding region were adjusted for the conditional analyses

7