Common variants in breast cancer risk loci predispose to distinct tumor subtypes

Thomas U. Ahearn[1*], Haoyu Zhang[1,2*], Kyriaki Michailidou[3,4,5], Roger L. Milne[6,7,8], Manjeet K. Bolla[3], Joe Dennis[3], Alison M. Dunning[9], Michael Lush[3], Qin Wang[3], Irene L. Andrulis[10], Hoda Anton-Culver[11], Volker Arndt[12], Kristan J. Aronson[13], Paul L. Auer[14,15], Annelie Augustinsson[16], Adinda Baten[17], Heiko Becher[18], Sabine Behrens[19], Javier Benitez[20], Marina Bermisheva[21], Carl Blomqvist[22,23], Stig E. Bojesen[24,25,26,27], Bernardo Bonanni[28], Anne-Lise Børresen-Dale[29,30], Hiltrud Brauch[31,32,33], Hermann Brenner[12,34,33], Angela Brooks-Wilson[35,36], Thomas Brüning[37], Barbara Burwinkel[38,39], Federico Canzian[40], Jose E. Castelao[41], Jenny Chang-Claude[19,42], Stephen J. Chanock[1], Georgia Chenevix-Trench[43], Christine L. Clarke[44], NBCS Collaborators[†], J.Margriet Collée[45], Angela Cox[46], Simon S. Cross[47], Kamila Czene[48], Mary B. Daly[49], Peter Devilee[50,51], Thilo Dörk[52], Miriam Dwek[53], Diana M. Eccles[54], D.Gareth Evans[55,56], Peter A. Fasching[57,58], Jonine Figueroa[59,60], Giuseppe Floris[17], Manuela Gago-Dominguez[61,62], Susan M. Gapstur[63], José A. García-Sáenz[64], Mia M. Gaudet[63], Graham G. Giles[6,7,8], Mark S. Goldberg[65,66,67], David E. Goldgar[68], Anna González-Neira[27], Grethe I. GrenakerAlnæs[29], Mervi Grip[69], Pascal Guénel[70], Christopher A. Haiman[71], Per Hall[48,72], Ute Torres[73], Elaine F. Harkness[74,75], Bernadette A.M. Heemskerk-Gerritsen[76], Bernd Holleczek[77], Antoinette Hollestelle[76], Maartje J. Hooning[76], Robert N. Hoover[1], John L. Hopper[7], Anthony Howell[78], kConFab/AOCS Investigators[†], Milena Jakimovska[79], Anna Jakubowska[80,81], Esther M. John[82], Michael E. Jones[83], Audrey Jung[19], Rudolf Kaaks[19], Saila Kauppila[84], Renske Keeman[85], Elza Khusnutdinova[86], Cari M. Kitahara[87], Yon-Dschun Ko[88], Stella Koutros[1], Vessela N. Kristensen[29,30], Ute Krüger[16], Katerina Kubelka-Sabit[89], Allison W. Kurian[82], Kyriacos Kyriacou[5,90], Diether Lambrechts[91,92], Derrick G. Lee[93,94], Annika Lindblom[95,96], Martha Linet[87], Jolanta Lissowska[97], Ana Llaneza[98], Wing-Yee Lo[31,99], Robert J.

23 MacInnis[6,7], Arto Mannermaa[100,101,102], Mehdi Manoochehri[73], Sara Margolin[72,103], Maria Elena
24 Martinez[62,104], Catriona McLean[105], Alfons Meindl[106], Usha Menon[107], Heli Nevanlinna[108],
25 William G. Newman[55,56], Jesse Nodora[109,110], Kenneth Offit[111], Håkan Olsson[16,112], Nick Orr[113],
26 Tjoung-Won Park-Simon[52], Julian Peto[3,114], Guillermo Pita[115], Dijana Plaseska-Karanfilska[79], Ross
27 Prentice[14], Kevin Punie[17], Katri Pylkäs[116,117], Paolo Radice[118], Gad Rennert[119], Atocha Romero[120],
28 Thomas Rüdiger[121], Emmanouil Saloustros[122], Sarah Sampson[123], Dale P. Sandler[124], Elinor J.
29 Sawyer[125], Rita K. Schmutzler[126], Minouk J. Schoemaker[83], Ben Schöttker[12], Mark E. Sherman[127],
30 Xiao-Ou Shu[128], Snezhana Smichkoska[129], Melissa C. Southey[8], John J. Spinelli[130,131], Anthony J.
31 Swerdlow[83,132], Rulla M. Tamimi[133,134,135], William J. Tapper[136], Jack A. Taylor[124,137], MaryBeth
32 Terry[138], Diana Torres[73,139], Melissa A. Troester[140], Celine M. Vachon[141], Carolien H.M. van
33 Deurzen[142], Elke M. van Veen[55,56], Philippe Wagner[16], Clarice R. Weinberg[143], Camilla Wendt[103],
34 Jelle Wesseling[85], Robert Winqvist[51,116,117,144], Alicja Wolk[81,145,146,81], Xiaohong R. Yang[1], Wei
35 Zheng[128], Fergus J. Couch[147], Jacques Simard[148], Peter Kraft[134,135], Douglas F. Easton[3,9], Paul D.P.
36 Pharoah[3,9], Marjanka K. Schmidt[85,149], Montserrat García-Closas[1**], Nilanjan Chatterjee[2,150**]

37

38 [1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of
39 Health, Department of Health and Human Services, Bethesda, MD, USA, [2]Department of
40 Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, [3]Centre for
41 Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of
42 Cambridge, Cambridge, UK, [4]Biostatistics Unit, The Cyprus Institute of Neurology and Genetics,
43 Nicosia, Cyprus, [5]Cyprus School of Molecular Medicine, Nicosia, Cyprus, [6]Cancer Epidemiology
44 Division, Cancer Council Victoria, Melbourne, Victoria, Australia, [7]Centre for Epidemiology and
45 Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne,
46 Melbourne, Victoria, Australia, [8]Precision Medicine, School of Clinical Sciences at Monash
47 Health, Monash University, Clayton, Victoria, Australia, [9]Centre for Cancer Genetic
48 Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK, [10]Fred A.
49 Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai
50 Hospital, Toronto, ON, Canada, [11]Department of Epidemiology, Genetic Epidemiology Research
51 Institute, University of California Irvine, Irvine, CA, USA, [12]Division of Clinical Epidemiology and
52 Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany, [13]Department
53 of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON,
54 Canada, [14]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA,
55 USA, [15]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA,
56 [16]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden,
57 [17]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute,
58 University Hospitals Leuven, Leuven, Belgium, [18]Institute of Medical Biometry and
59 Epidemiology, University of Hamburg, Hamburg, Germany, [19]Division of Cancer Epidemiology,

60 German Cancer Research Center (DKFZ), Heidelberg, Germany, [20]Centro de Investigación en
61 Red de Enfermedades Raras (CIBERER), Valencia, Spain, [21]Institute of Biochemistry and
62 Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa, Russia, [22]Department of
63 Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, [23]Department of
64 Oncology, Örebro University Hospital, Örebro, Sweden, [24]Copenhagen General Population
65 Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark,
66 [25]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University
67 Hospital, Herlev, Denmark, [26]Faculty of Health and Medical Sciences, University of Copenhagen,
68 Copenhagen, Denmark, [27]Human Cancer Genetics Programme, Spanish National Cancer
69 Research Centre (CNIO), Madrid, Spain, [28]Division of Cancer Prevention and Genetics, IEO,
70 European Institute of Oncology IRCCS, Milan, Italy, [29]Department of Cancer Genetics, Institute
71 for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway, [30]Institute of
72 Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, [31]Dr. Margarete
73 Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany, [32]iFIT-Cluster of
74 Excellence, University of Tübingen, Tübingen, Germany, [33]German Cancer Consortium (DKTK),
75 German Cancer Research Center (DKFZ), Heidelberg, Germany, [34]Division of Preventive
76 Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases
77 (NCT), Heidelberg, Germany, [35]Genome Sciences Centre, BC Cancer Agency, Vancouver, BC,
78 Canada, [36]Department of Biomedical Physiology and Kinesiology, Simon Fraser University,
79 Burnaby, BC, Canada, [37]Institute for Prevention and Occupational Medicine of the German
80 Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany,
81 [38]Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg,

82 Germany, [39]Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg,
83 University of Heidelberg, Heidelberg, Germany, [40]Genomic Epidemiology Group, German
84 Cancer Research Center (DKFZ), Heidelberg, Germany, [41]Oncology and Genetics Unit, Instituto
85 de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS,
86 Vigo, Spain, [42]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH),
87 University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [43]Department of Genetics
88 and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland,
89 Australia, [44]Westmead Institute for Medical Research, University of Sydney, Sydney, New South
90 Wales, Australia, [45]Department of Clinical Genetics, Erasmus University Medical Center,
91 Rotterdam, The Netherlands, [46]Sheffield Institute for Nucleic Acids (SInFoNiA), and Weston Park
92 Cancer Centre, Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK,
93 [47]Department of Neuroscience, University of Sheffield, Sheffield, UK, [48]Department of Medical
94 Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [49]Department of
95 Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA, [50]Department of Pathology,
96 Leiden University Medical Center, Leiden, The Netherlands, [51]Department of Human Genetics,
97 Leiden University Medical Center, Leiden, The Netherlands, [52]Gynaecology Research Unit,
98 Hannover Medical School, Hannover, Germany, [53]Department of Biomedical Sciences, Faculty
99 of Science and Technology, University of Westminster, London, UK, [54]Cancer Sciences Academic
100 Unit, Faculty of Medicine, University of Southampton, Southampton, UK, [55]Division of Evolution
101 and Genomic Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health,
102 University of Manchester, Manchester Academic Health Science Centre, Manchester, UK,
103 [56]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester NIHR Biomedical

104 Research Centre, Manchester University Hospitals NHS, Foundation Trust, Manchester
105 Academic Health Science Centre, Manchester, UK, [57]David Geffen School of Medicine,
106 Department of Medicine Division of Hematology and Oncology, University of California at Los
107 Angeles, Los Angeles, CA, USA, [58]Department of Gynecology and Obstetrics, Comprehensive
108 Cancer Center ER-EMN, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-
109 Nuremberg, Erlangen, Germany, [59]Usher Institute of Population Health Sciences and
110 Informatics, The University of Edinburgh Medical School, Edinburgh, UK, [60]Cancer Research UK
111 Edinburgh Centre, Edinburgh, UK, [61]Genomic Medicine Group, Galician Foundation of Genomic
112 Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo
113 Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain, [62]Moores
114 Cancer Center, University of California San Diego, La Jolla, CA, USA, [63]Behavioral and
115 Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA, [64]Medical Oncology
116 Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC),
117 Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain, [65]Department of
118 Medicine, McGill University, Montréal, QC, Canada, [66]Division of Clinical Epidemiology, Royal
119 Victoria Hospital, McGill University, Montréal, QC, Canada, [67]Breast Cancer Research Unit,
120 Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia,
121 [68]Department of Dermatology, Huntsman Cancer Institute, University of Utah School of
122 Medicine, Salt Lake City, UT, USA, [69]Department of Surgery, Oulu University Hospital, University
123 of Oulu, Oulu, Finland, [70]Cancer & Environment Group, Center for Research in Epidemiology
124 and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif,
125 France, [71]Department of Preventive Medicine, Keck School of Medicine, University of Southern

126 California, Los Angeles, CA, USA, [72]Department of Oncology, Södersjukhuset, Stockholm,
127 Sweden, [73]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ),
128 Heidelberg, Germany, [74]Division of Informatics, Imaging and Data Sciences, Faculty of Biology,
129 Medicine and Health, University of Manchester, Manchester Academic Health Science Centre,
130 Manchester, UK, [75]Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester
131 University NHS Foundation Trust, Manchester, UK, [76]Department of Medical Oncology, Family
132 Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands, [77]Saarland Cancer
133 Registry, Saarbrücken, Germany, [78]Division of Cancer Sciences, University of Manchester,
134 Manchester, UK, [79]Research Centre for Genetic Engineering and Biotechnology "Georgi D.
135 Efremov", MASA, Skopje, Republic of North Macedonia, [80]Department of Genetics and
136 Pathology, Pomeranian Medical University, Szczecin, Poland, [81]Independent Laboratory of
137 Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland,
138 [82]Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University
139 School of Medicine, Stanford, CA, USA, [83]Division of Genetics and Epidemiology, The Institute of
140 Cancer Research, London, UK, [84]Department of Pathology, Oulu University Hospital, University
141 of Oulu, Oulu, Finland, [85]Division of Molecular Pathology, The Netherlands Cancer Institute -
142 Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, [86]Department of Genetics and
143 Fundamental Medicine, Bashkir State University, Ufa, Russia, [87]Radiation Epidemiology Branch,
144 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA,
145 [88]Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter
146 Krankenhaus, Bonn, Germany, [89]Department of Histopathology and Cytology, Clinical Hospital
147 Acibadem Sistina, Skopje, Republic of North Macedonia, [90]Department of Electron

148 Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia,
149 Cyprus, [91]VIB Center for Cancer Biology, VIB, Leuven, Belgium, [92]Laboratory for Translational
150 Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium,
151 [93]Department of Mathematics and Statistics, St. Francis Xavier University, Antigonish, Canada,
152 [94]Cancer Control Research, BC Cancer Agency, Vancouver, BC, Canada, [95]Department of
153 Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden, [96]Department of
154 Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden, [97]Department of Cancer
155 Epidemiology and Prevention, M. Sklodowska-Curie Cancer Center, Oncology Institute, Warsaw,
156 Poland, [98]General and Gastroenterology Surgery Service, Hospital Universitario Central de
157 Asturias, Oviedo, Spain, [99]University of Tübingen, Tübingen, Germany, [100]Translational Cancer
158 Research Area, University of Eastern Finland, Kuopio, Finland, [101]Institute of Clinical Medicine,
159 Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland, [102]Imaging
160 Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland,
161 [103]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet,
162 Stockholm, Sweden, [104]Department of Family Medicine and Public Health, University of
163 California San Diego, La Jolla, CA, USA, [105]Department of Anatomical Pathology, The Alfred
164 Hospital, Melbourne, Victoria, Australia, [106]Department of Gynecology and Obstetrics, Ludwig
165 Maximilian University of Munich, Munich, Germany, [107]MRC Clinical Trials Unit at UCL, Institute
166 of Clinical Trials & Methodology, University College London, London, UK, [108]Department of
167 Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland,
168 [109]Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA, [110]Department
169 of Family Medicine and Public Health, School of Medicine, University of California, San Diego, La

170 Jolla, CA, USA, [111]Clinical Genetics Research Lab, Department of Cancer Biology and Genetics,
171 Memorial Sloan-Kettering Cancer Center, New York, NY, USA, [112]Clinical Genetics Service,
172 Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA,
173 [113]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, Ireland, UK,
174 [114]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and
175 Tropical Medicine, London, UK, [115]Human Genotyping-CEGEN Unit, Human Cancer Genetic
176 Program, Spanish National Cancer Research Centre, Madrid, Spain, [116]Laboratory of Cancer
177 Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu,
178 University of Oulu, Oulu, Finland, [117]Laboratory of Cancer Genetics and Tumor Biology,
179 Northern Finland Laboratory Centre Oulu, Oulu, Finland, [118]Unit of Molecular Bases of Genetic
180 Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei
181 Tumori (INT), Milan, Italy, [119]Clalit National Cancer Control Center, Carmel Medical Center and
182 Technion Faculty of Medicine, Haifa, Israel, [120]Medical Oncology Department, Hospital
183 Universitario Puerta de Hierro, Madrid, Spain, [121]Institute of Pathology, Staedtisches Klinikum
184 Karlsruhe, Karlsruhe, Germany, [122]Department of Oncology, University Hospital of Larissa,
185 Larissa, Greece, [123]Prevent Breast Cancer Centre and Nightingale Breast Screening Centre,
186 Manchester University NHS Foundation Trust, Manchester, UK, [124]Epidemiology Branch,
187 National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA,
188 [125]Research Oncology, Guy’s Hospital, King's College London, London, UK, [126]Center for
189 Hereditary Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne,
190 University of Cologne, Cologne, Germany, [127]Department of Health Sciences Research, Mayo
191 Clinic College of Medicine, Jacksonville, FL, USA, [128]Division of Epidemiology, Department of

192 Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt
193 University School of Medicine, Nashville, TN, USA, [129]Ss. Cyril and Methodius University in
194 Skopje, Medical Faculty, University Clinic of Radiotherapy and Oncology, Skopje, Republic of
195 North Macedonia, [130]Population Oncology, BC Cancer, Vancouver, BC, Canada, [131]School of
196 Population and Public Health, University of British Columbia, Vancouver, BC, Canada, [132]Division
197 of Breast Cancer Research, The Institute of Cancer Research, London, UK, [133]Channing Division
198 of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard
199 Medical School, Boston, MA, USA, [134]Department of Epidemiology, Harvard T.H. Chan School of
200 Public Health, Boston, MA, USA, [135]Program in Genetic Epidemiology and Statistical Genetics,
201 Harvard T.H. Chan School of Public Health, Boston, MA, USA, [136]Faculty of Medicine, University
202 of Southampton, Southampton, UK, [137]Epigenetic and Stem Cell Biology Laboratory, National
203 Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA,
204 [138]Department of Epidemiology, Mailman School of Public Health, Columbia University, New
205 York, NY, USA, [139]Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota,
206 Colombia, [140]Department of Epidemiology, Gilliungs School of Global Public Health and UNC
207 Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel
208 Hill, NC, USA, [141]Department of Health Science Research, Division of Epidemiology, Mayo Clinic,
209 Rochester, MN, USA, [142]Department of Pathology, Erasmus University Medical Center,
210 Rotterdam, The Netherlands, [143]Biostatistics and Computational Biology Branch, National
211 Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA,
212 [144]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, [145]Institute
213 of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, [146]Department of

214 Surgical Sciences, Uppsala University, Uppsala, Sweden, [147]Department of Laboratory Medicine
215 and Pathology, Mayo Clinic, Rochester, MN, USA, [148]Genomics Center, Centre Hospitalier
216 Universitaire de Québec – Université Laval, Research Center, Québec City, QC, Canada,
217 [149]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute -
218 Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands, [150]Department of Oncology,
219 School of Medicine, Johns Hopkins University, Baltimore, MD, USA

220

221 *These authors contributed equally to this work

222 **These authors contributed equally to this work

223 †Lists of participants and their affiliations appear in the Supplementary Information

224 **Conflicts of interest:** None to report

225 **Corresponding Author**

226 Montserrat Garcia-Closas

227 National Cancer Institute

228 Division of Cancer Epidemiology & Genetics

229 9609 Medical Center Drive

230 Rockville, MD 20850

231 montserrat.garcia-closas@nih.gov

232

Abstract

Genome-wide association studies (GWAS) have identified over 170 common breast cancer susceptibility variants using standard GWAS methods. Many of these variants have differential associations by estrogen receptor (ER), but how these variants relate with other tumor features and intrinsic molecular subtypes is unclear. We used two-stage polytomous logistic regression models to evaluate variants in relation to multiple tumor features (ER, progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and grade) adjusting for each other, and to intrinsic-like subtypes. Eighty-five of 178 variants were associated with at least one tumor feature (false discovery rate <5%), most commonly ER and grade followed by PR and HER2. Case-control comparisons among these 85 variants identified 65 variants strongly or exclusively associated (P<0.05) with luminal-like subtypes, 5 variants associated with all subtypes at differing strengths and 15 variants primarily associated with non-luminal subtypes, especially triple-negative (TN) disease. Five variants were associated with risk of Luminal A-like and TN subtypes in opposite directions. This report demonstrates a high level of complexity in the etiology heterogeneity of breast cancer susceptibility loci and can inform investigations of subtype-specific risk prediction.

249 ## Introduction

250 Breast cancer represents a heterogenous group of diseases with different molecular and
251 clinical features[1]. Clinical assessment of estrogen receptor (ER), progesterone receptor (PR),
252 human epidermal growth factor receptor 2 (HER2) and histological grade are routinely
253 determined to inform treatment strategies and prognostication.[2] Combined, these tumor
254 features define five intrinsic-like subtypes (i.e. luminal A-like, luminal B/HER2-negative-like,
255 luminal B-like, HER2-enriched-like, and basal-like/triple negative) that are correlated with
256 intrinsic subtypes defined by gene expression panels[2]. Most known breast cancer risk or
257 protective factors are related to luminal or hormone receptor (ER or PR; HR) positive tumors,
258 whereas less is known about the etiology of triple-negative (TN) tumors, an aggressive subtype[3,
259 4].

260 Breast cancer genome-wide association studies (GWAS) have identified over 170
261 common susceptibility variants, most of them single nucleotide polymorphisms (SNPs), of which
262 many are differentially associated with ER-positive than ER-negative disease[5, 6, 7]. These include
263 20 variants that primarily predispose to ER-negative or TN disease[6, 7]. However, few studies
264 have evaluated variant associations with other tumor features, or simultaneously studied
265 multiple, correlated tumor markers to identify source(s) of etiologic heterogeneity[6, 8, 9, 10, 11, 12].
266 We recently developed a two-stage polytomous logistic regression method that efficiently
267 characterizes etiologic heterogeneity while accounting for tumor marker correlations and
268 missing tumor data[13, 14]. This method can help describe complex relationships between
269 susceptibility variants and multiple tumor features, helping to clarify breast cancer subtype
270 etiologies and increasing the power to generate more accurate risk estimates between

susceptibility variants and less common subtypes. We recently demonstrated the power of this method in a GWAS to identify novel breast cancer susceptibility accounting for tumor heterogeneity[14].

In this report, we sought to expand our understanding of etiologic heterogeneity among breast cancer subtypes, by applying the two-stage polytomous logistic regression methodology to a large study population from the Breast Cancer Association Consortium (BCAC) for detailed characterization of risk associations of 178 known risk variants by tumor subtypes defined by ER, PR, HER2 and tumor grade.

Results

The mean (SD) ages at diagnosis (cases) and enrollment (controls) were 56.6 (12.2) and 56.4 (12.2) years, respectively. Eighty-one percent of tumors were ER-positive, 68% PR-positive, 83% HER2-negative and 69% grade 1 or 2 (**Table 1; Supplementary Table 1**). The most common intrinsic-like subtype was luminal A-like (59%), followed by TN (13%), luminal B/HER2-negative-like (12%), Luminal B-like (12%) and HER2-enriched-like (5%; **Table 1**).

Global test for heterogeneity (primary hypothesis)

The two-stage models fitted for each of the 178 variants included terms for ER, PR, HER2 and grade, and identified 85 of 178 variants (47.7%) with evidence for heterogeneity by at least one tumor feature (FDR<5%; **Figure 1 and 2; Supplementary Fig. 1**).

Description of sources of heterogeneity by each tumor feature

Marker-specific analyses identified ER and grade as the two features that most often

293 contributed to the observed heterogeneity (45 and 34 variants respectively had case-case
294 marker-specific P<0.05), and 30 variants were associated with more than one tumor
295 characteristic (**Figure 1 and 2; Supplementary Fig. 1)**. Seventeen of these 85 variants showed
296 no associations in case-case comparisons with any individual tumor marker at P<0.05 in their
297 corresponding fixed-effect two-stage models (**Supplementary Fig. 1**). Twenty-one variants
298 showed evidence for having ER as the only source of heterogeneity (**Supplementary Fig. 1**
299 **footnote 'm'**), 12 variants showed evidence for having grade as the only source of
300 heterogeneity (**Supplementary Fig. 1 footnote 'l'**), four variants showed evidence for PR as the
301 only source of heterogeneity (**Supplementary Fig. 1 footnote 'n'**) and one variant showed
302 evidence for HER2 as the only source of heterogeneity (**Supplementary Fig. 1 footnote 'o'**).
303
304 *Description of variant associations with intrinsic-like molecular tumor subtypes*
305 Case-control comparisons for the 85 variants with evidence for global heterogeneity
306 identified four main clusters of variants according to the p-values for risk associations with each
307 subtype (**Figure 3 and Supplementary Fig. 2**). Sixty-five variants in cluster 1 (n=3 variants) and
308 cluster 4 (n=62 variants) showed the strongest evidence for associations with risk for luminal-
309 like subtypes, cluster 2 (n=5 variants) was associated with risk for all subtypes at varying
310 strengths and cluster 3 (n=15 variants) with stronger evidence for TN or non-luminal subtype
311 associations. Five variants (rs67397200, rs6678914 and rs4577244 in cluster 3 and rs7072776
312 and rs17879961 in cluster 4) were found associated in opposite directions with luminal A-like
313 and TN subtypes. **Supplementary Table 2** shows the case-control associations between all 178
314 variants and the intrinsic-like subtypes.

315 Cluster 1 included two correlated (r^2=0.73) variants at 10q26.13, rs2981578 and
316 rs35054928, that were strongly associated in case-control analyses with risk of all the luminal-
317 like subtypes (e.g. OR (95%CI)=1.29 (1.27 to 1.31), P=2.01x10^{-231} and OR (95%CI)=1.35 (1.33 to
318 1.37), P=5.00x10^{-300} for luminal A-like, respectively), weakly associated with risk of HER2-
319 enriched-like subtype (OR (95%CI)=1.11 (1.05 to1.16), P=6.32x10^{-05} and OR (95% CI)=1.10 (1.04
320 to 1.15), P=3.07x10^{-4}, respectively), but not associated (P>0.05) with risk of TN disease (**Figures
321 3-4 and Supplementary Fig. 2**). In case-case comparisons both rs2981578 and rs35054928
322 showed the strongest evidence for associations with ER (**Supplementary Fig. 1**). In the
323 extended two-stage model that additionally included a TN term to test for heterogeneity
324 between TN vs non-TN subtypes both rs2981578 and rs35054928 were associated with TN
325 status (**Supplementary Fig. 3**). A third variant in 10q26.13, rs45631563, showed similar
326 association patterns, but fell into cluster 4 since, unlike rs2981578 and rs35054928, it was not
327 associated (P>0.05) in case-control analyses with risk of the HER2-enriched-like subtype (**Figure
328 3 and Supplementary Figs. S2, S4**). Case-case comparisons for rs45631563 showed associations
329 with ER and grade. A third variant in cluster 1, 16q12.1-rs4784227, was most strongly
330 associated with risk of luminal-like subtypes (e.g. OR (95%CI)=1.28 (1.26 to 1.30), P=6.68x10^{-176}
331 for luminal A-like), and the HER2-enriched-like tumors (OR (95%CI)=1.26 (1.19 to 1.33),
332 P=3.04x10^{-16}) and weaker so with TN disease (OR (95%CI)=1.11 (1.08 to 1.15), P=6.05x10^{-10};
333 **Figures 3-4 and, Supplementary Figs. 2**). In case-case analyses, rs4784227 was associated with
334 all tumor markers, particularly PR (**Supplementary Fig. 1**).
335 Five variants in cluster 2 were associated in case-control comparisons, to different
336 extents, with risk of all five intrinsic-like subtypes. 6q25-rs2747652 and 1p36.22-rs616488 were

337 associated particularly with risk of HER2-positive subtypes (**Figures 3-4**; **Supplementary Fig. 2**),
338 and in case-case comparisons these variants were associated with HER2 status and grade
339 (**Supplementary Fig. 1**). Two additional variants in 6q25 showed the strongest evidence for
340 being associated with risk of TN disease (OR (95%CI)=1.30 (1.24 to 1.38), P=8.26x10^{-23} for
341 rs9397437 and OR (95%CI)=1.15 (1.12 to 1.19), P=1.24x10^{-19} for rs3757322; **Figures 3-4**;
342 **Supplementary Fig. 2**), and in case-case comparisons they were associated with ER and grade
343 (**Supplementary Fig. 1**). 13q13.1-rs11571833 was associated in case-control comparisons with
344 risk of all subtypes, but case-case comparisons showed an association only with grade
345 (**Supplementary Fig. 1**). In case-control comparisons the ORs (95% CIs) for rs11571833 with
346 grade 3, grade 2 and grade 1 subtypes were: 1.48 (1.36 to 1.62), P=2.3x10^{-19}; 1.27 (1.18 to
347 1.35), P=5.0x10^{-12}; and 1.08 (0.97-1.20), P=0.15, respectively (**Supplementary Fig. 5**).
348 Cluster 3 included 15 variants most strongly associated in case-control comparisons with
349 risk of HR-negative subtypes, including three variants with the strongest evidence for
350 associations with risk of TN disease: 19p13.11-rs67397200 (OR (95% CI)=1.27 (1.23 to 1.31),
351 P=1.07x10^{-50}), 5p15.33-rs10069690 (OR (95% CI)=1.27 (1.23 to 1.31), P=3.79x10^{-48}) and 1q32.11-
352 rs4245739 (OR (95% CI)=1.18 (1.14 to 1.22), P=2.72x10^{-23}). Two variants at 11q22.3, rs11374964
353 (OR (95% CI)=0.90 (0.88-0.93), P=2.71x10^{-11}) and rs74911261 (OR (95% CI)=0.93 (0.90-0.96),
354 P=2.71x10^{-11}), were associated (P<0.05) only with TN disease (**Figures 3-4; Supplementary Fig.**
355 **2**). In the extended model case-case comparisons these five variants were associated with TN
356 status (**Supplementary Fig. 3**). The remaining 10 variants in case-control comparisons were all
357 associated with risk of TN disease (P<0.05; **Figures 3-4; Supplementary Fig. 3**), and in case-case
358 comparisons seven of these 10 variants showed the evidence for ER as the only source of

359 heterogeneity. 5p15.33-rs3215401 was associated with both ER and PR, and 2p24.1-
360 rs12710696 and 19q12-rs113701136 showed no evidence in case-case comparisons for
361 associations with any of the individual tumor markers (**Supplementary Fig. 1**). Besides
362 rs11374964 and rs74911261, variants in this cluster were not HR-negative or TN-specific
363 variants, as they were also associated in case-control comparisons with risk of luminal-like
364 subtypes. Three variants showed weak associations with luminal A-like disease (OR (95%
365 CI)=0.98 (0.97 to 1.00); P=0.039 for rs67397200; OR (95% CI)=1.02 (1.00 to 1.03); P=0.024 for
366 rs6678914; and OR (95% CI)=1.02 (1.00 to 1.04); P=0.02 for rs4577244) in an opposite direction
367 to their associations with TN disease (OR (95% CI)= 1.27 (1.23 to 1.31); P=1.07×10^{-50} for
368 rs67397200; OR (95% CI)=0.93 (0.91 to 0.96); P=1.07×10^{-4} for rs6678914 and OR (95% CI)=0.90
369 (0.86 to 0.93); P=2.99×10^{-9} for rs4577244; **Figures 3-4**; **Supplementary Fig. 3**).
370 Cluster 4 (n=62 variants) showed evidence in case-control comparisons for associations
371 with risk of luminal-subtypes, especially luminal A-like disease. The five variants in this cluster
372 showing the strongest evidence for associations with risk of luminal A-like disease were:
373 5q11.2-rs62355902 (OR (95% CI)=1.22 (1.20 to 1.25), P=3.94×10^{-85}), 11q13.3-rs75915166 (OR
374 (95% CI)=1.44 (1.40 to 1.48), P=2.26×10^{-131}), 11q13.3-rs554219 (OR (95% CI)=1.33 (1.30 to 1.37),
375 P=1.31×10^{-98}), 1p11.2-rs11249433 (OR (95% CI)=1.17 (1.15 to 1.19), P=5.25×10^{-90}) and 5p12-
376 rs10941679 (OR (95% CI)=1.20 (1.18 to 1.22), P=1.39×10^{-93}; **Figure 3 and Figures S2, S4**). In case-
377 case comparisons these five variants were associated with grade and with at least one other
378 tumor marker (**Supplementary Fig. 1**). Eighteen of the 62 variants also showed weaker
379 evidence in case-control comparisons for associations with non-luminal subtypes, 11 of which
380 were associated with risk of TN disease (P<0.05; **Figure 3 and Supplementary Figs. S2, S4**).

381 Notably, rs7072776 and rs17879961 (I157T) were associated in opposite directions with risk of
382 luminal A-like disease (OR (95% CI)=1.10 (1.08 to 1.11); P=4.96x10^{-27} for rs7072776; and OR
383 (95% CI)=1.44 (1.31 to 1.59); P=9.26x10^{-14} for rs17879961) and risk of TN disease (OR (95% CI)=
384 0.96 (0.93 to 0.99); P=0.02 for rs7072776 and OR (95% CI)=0.61 (0.47 to 0.80); P=2.55x10^{-4} for
385 rs17879961). In case-case comparisons rs7072776 was associated with ER and grade
386 (**Supplementary Fig. 1**), and rs17879961 was associated with the TN phenotype in the extended
387 model (**Supplementary Fig. 3**).
388 **Supplementary Fig. 5** shows case-control associations by tumor grade for the 12
389 variants associated exclusively with grade in case-case comparisons. rs11571833, rs17426269
390 and rs11820646 showed stronger evidence for predisposing to risk of high-grade subtypes, and
391 the remaining variants showed stronger evidence for predisposing to risk of low-grade
392 subtypes.
393

394 **Discussion**

395 This study demonstrates the extent and complexity of genetic etiologic heterogeneity
396 among known common breast cancer susceptibility variants by multiple tumor characteristics,
397 using novel methodology in the largest and the most comprehensive investigation conducted to
398 date. We found compelling evidence that about half of the investigated breast cancer
399 susceptibility loci (85 of 178 variants) predispose to tumors of different characteristics. We
400 identified tumor grade, along with confirming ER and TN status, as important determinants of
401 etiologic heterogeneity. Associations with individual tumor features translated into differential
402 associations with the risk of intrinsic-like subtypes defined by their combinations.

403 Many of the variants showing subtype heterogeneity predisposed to risk of all subtypes,
404 but with different magnitudes. For example, 21 of 65 variants found predominantly associated
405 with risk of luminal-like subtypes were also associated with risk of at least one of the non-
406 luminal subtypes. These include variants identified in early GWAS for overall breast cancer, such
407 as variants in the loci for *FGFR2* (rs35054928 and rs2981578)[15, 16] and 8q24.21 (rs13281615)[15]
408 that were associated with luminal-like and HER2-enriched-like subtypes. rs4784227 located
409 near *TOX3*[15, 17] and rs62355902 located in a *MAP3K1*[15] regulatory element, were associated
410 with risk of all five subtypes. Of the five variants found associated in opposite directions with
411 luminal A-like and TN disease, we previously reported rs6678914 and rs4577244 to have
412 opposite effects between ER-negative and ER-positive tumors[6]. rs17879961 (I157T), a likely
413 causal[18] mis-sense variant located in a *CHEK2* functional domain that reduces or abolishes
414 substrate binding[19], was previously reported to have opposite directions of effects on lung
415 adenocarcinoma and lung squamous cell carcinoma and for lung cancer between smokers and
416 non-smokers.[20, 21] However, further studies are required to follow-up and clarify the
417 mechanisms for these apparent cross-over effects.

418 In prior ER-negative GWAS we identified 20 variants that predispose to ER-negative
419 disease, of which five variants were only or most strongly associated with risk of TN disease
420 (rs4245739, rs10069690, rs74911261, rs11374964, and rs67397200)[6, 7]. We further confirmed
421 these five variants to be most strongly associated with TN disease and that they were
422 associated in case-case comparisons with TN status in the extended model. The remaining
423 previously identified 15 variants all showed associations with risk of HR-negative disease, and
424 for all but four variants (rs17350191, rs200648189, rs6569648, and rs322144) evidence of

425 global heterogeneity was observed. Among the variants in cluster 3, rs3215401 was the only
426 variant that was not identified in a prior ER-negative GWAS[6, 7] rs3215401 was identified in a
427 fine-mapping analysis of *TERT* and, consistent with our findings, reported to be most strongly
428 associated with risk of ER-negative disease but was also associated with risk of ER-positive
429 disease[22].

430 Little is known regarding PR and HER2 as sources of etiologic heterogeneity independent
431 of ER or TN status. Of the four variants that showed evidence for heterogeneity only according
432 to PR, rs10759243[5, 23], rs11199914[5] and rs72749841[5] were previously found primarily
433 associated with risk of ER-positive disease, and rs10816625 was found to be associated with risk
434 of ER-positive/PR-positive tumors, but not other ER/PR combinations[11]. rs10995201 was the
435 only variant found in case-case comparisons to be solely associated with HER2 status, although
436 the evidence was not strong, requiring further confirmation. Previously rs10995201 showed no
437 evidence of being associated with ER status[24]. Among all variants found with PR or HER2
438 associations, few have been investigated for PR or HER2 heterogeneity while adjusting for ER[8, 9,
439 10, 11, 12]. We previously reported rs10941679 to be associated with PR-status, independent of ER,
440 and also with grade[9]. We also found suggestive evidence of PR-specific heterogeneity for
441 16q12-rs3803662[12], which is in high LD (r^2= 0.78) with rs4784227 (*TOX3*), which was strongly
442 associated with PR status. Our findings for rs2747652 are also consistent with a prior BCAC fine-
443 mapping analysis across the *ESR1* locus, which found rs2747652 to be associated with risk of
444 the HER2-enriched subtype and high grade independent of ER[8]. rs2747652 overlaps an
445 enhancer region and is associated with reduced *ESR1* and *CCDC170* expression[8].

446 Histologic grade is a composite of multiple tumor characteristics including mitotic count,
447 nuclear pleomorphism, and degree of tubule or gland formation[25]. Among the 12 variants
448 identified with evidence of heterogeneity by grade only, rs17426269, rs11820646, and
449 rs11571833 were found most strongly associated with risk of grade 3 disease. rs11571833 lies
450 in the *BRCA2* coding region and produces a truncated form of the protein[26] and has been shown
451 to be associated with both risk of TN disease and risk of serous ovarian tumors, both of which
452 tend to be high-grade[27]. To our knowledge, rs17426269 and rs11820646 have not been
453 investigated in relation to grade heterogeneity. The remaining 9 variants were all more strongly
454 associated with grade 1 or grade 2 disease. Five of these variants were previously reported to
455 be associated primarily with ER-positive disease[5, 28, 29], highlighting the importance of
456 accounting for multiple tumor characteristics to better illuminate heterogeneity sources.

457 For descriptive purposes, we performed clustered analyses that allowed visualization of
458 common heterogeneity patterns across variants. We identified 17 variants with evidence of
459 global heterogeneity (FDR<0.05) but no significant (marker-specific P<0.05) associations with
460 any of the individual tumor characteristic(s). This is likely explained by the fact that the test for
461 association with specific tumor markers using fixed-effects models are less powerful than
462 mixed-effects models used to test the primary hypothesis of global heterogeneity by any tumor
463 marker[13].

464 A major strength of our study is our large sample size of over 100,000 breast cancer
465 cases with tumor marker information, and a similar number of controls, making this the largest,
466 most comprehensive breast cancer heterogeneity investigation. Our application of the two-
467 stage polytomous logistic regression enabled adjusting for multiple, correlated tumor markers

468 and accounting for missing tumor marker data. This is a more powerful and efficient modeling
469 strategy for identifying heterogeneity sources among highly correlated tumor markers,
470 compared with standard polytomous logistic regression[13, 14]. In simulated and real data
471 analyses, we have demonstrated that in the presence of heterogenous associations across
472 subtypes, the two-stage model is more powerful than polytomous logistic regression for
473 detecting heterogeneity. Moreover, we have demonstrated that in the presence of correlated
474 markers, the two-stage model, incorporating all markers simultaneously, has much better
475 ability to distinguish the true source(s) of heterogeneity compared to testing for heterogeneity
476 by analysis of one marker at a time[13]. In prior analyses, we showed that the two-stage
477 polytomous regression is a powerful approach to identify susceptibility variants that display
478 tumor heterogeneity[14].

479 Our study also had limitations. First, many breast cancer cases from studies included in
480 this report had missing information on one or more tumor characteristics. To address this
481 limitation, we implemented an EM algorithm that allowed a powerful analysis to incorporate
482 cases with missing tumor characteristics under the assumption that tumor characteristics are
483 *missing at random* (MAR), i.e. the underlying reason for missing data may depend on observed
484 tumor markers or/and covariate values, but not on the missing values themselves[30]. If this
485 assumption is violated it can lead to an inflated type-one error[13]. However, in the context of
486 genetic association testing, the missingness mechanism would also need related to the genetic
487 variants under study, which is unlikely. Our study focused on investigating ER, PR, HER2, and
488 grade as heterogeneity sources and future studies with more detailed tumor characterization
489 could reveal additional etiologic heterogeneity sources.

490 In summary, our findings provide insights into the complex etiologic heterogeneity
491 patterns of common breast cancer susceptibility loci. These findings may inform future studies,
492 such as fine-mapping and functional analyses to identify the underlying causal variants,
493 clarifying biological mechanisms that drive genetic predisposition to breast cancer subtypes.
494 Moreover, these analyses provide precise estimates of relative risk for different intrinsic-like
495 subtypes that could improve the discriminatory accuracy of subtype-specific polygenic risk
496 scores[31].

497 ## Methods

498 **Study Population and Genotyping**

499 The study population and genotyping are described in previous publications[5, 6] and in
500 the **Supplemental Methods**. We included invasive cases and controls from 81 BCAC studies
501 with genotyping data from two Illumina genome-wide custom arrays, the iCOGS and
502 OncoArray (106,571 cases (OncoArray: 71,788; iCOGS: 34,783) and 95,762 controls
503 (OncoArray: 58,134; iCOGS: 37,628); **Supplementary Table 1**). We evaluated 178 breast
504 cancer susceptibility variants that were identified in or replicated by prior BCAC analyses. The
505 majority of these variants (n=158) were identified because of their association with risk of
506 overall breast cancer, and a small number of variants (n=20) were identified because of their
507 association specific to ER-negative breast cancer (**Supplementary Table 3**)[5, 6]. Genotypes for
508 the variants marking the 178 susceptibility loci were determined by genotyping with the
509 iCOGS and the OncoArray arrays and imputation to the 1000 Genomes Project (Phase 3)
510 reference panel.

511

512 **Statistical Analysis**

513 An overview of the analytic strategy is shown in **Figure** 1 and a detailed discussion of the
514 statistical methods, including of the two-stage polytomous logistic regression, are provided in
515 the **Supplementary Methods** and elsewhere[13, 14]. Briefly, our main aim was to identify which of
516 178 known susceptibility variants showed heterogenous associations by ER, PR, HER2 and
517 grade. Specifically, for each variant, the global null hypothesis was that there was no difference
518 in risk of breast cancer associated with the variant genotype across any of the tumor features
519 being evaluated.

520 We used a two-stage polytomous regression model to perform global test for
521 heterogeneity. The two-stage polytomous model allows modelling genetic association of breast
522 cancer accounting for possible heterogeneity in effects by combination of multiple tumor
523 markers using a parsimonious decomposition of subtype-specific case-control odds-ratio
524 parameters in terms of marker-specific case-case odd-ratio parameters. We introduced further
525 parsimony by using mixed-effect formulation of the model that allows ER-specific case-case
526 parameters to be treated as fixed and similar parameters for other markers (PR, HER2 and
527 grade) as random. We used an EM algorithm[32] for parameter estimation under this model to
528 account for missing data in tumor characteristics. A detailed description of the proposed
529 method and assessment of performance in simulation and real studies are presented in the
530 **Supplementary Methods** and as a separate manuscript[13].

531 In the main analysis, we accounted for multiple testing (178 tests, one for each of
532 variant) of the global heterogeneity test using a false discovery rate (FDR) <0.05 under the

533 Benjamini-Hochberg procedure[33]. For the variants showing evidence for global heterogeneity
534 after FDR adjustment, we further evaluated which of the tumor features contributed to the
535 heterogeneity by fitting a fixed-effects two-stage model that simultaneously tested for
536 associations with each tumor feature (this model was fitted for each variant separately). We
537 used a threshold of $P<0.05$ for marker-specific tumor heterogeneity tests to describe which
538 specific tumor marker(s) contributed to the observed heterogeneity, adjusting for the other
539 tumor markers in the model. This p-value threshold was used only for descriptive purposes, as
540 the primary hypotheses were tested using the FDR-adjusted global test for heterogeneity
541 described above.

542 We conducted various additional analyses to explore for evidence of heterogeneity. We
543 fitted an extended two-stage model with an additional term for TN status to test for differences
544 between TN vs non-TN subtypes. We also fitted a two-stage model to estimate case-control ORs
545 and 95% confidence intervals (CI) between the variants and five intrinsic-like subtypes defined
546 by combinations of ER, PR, HER2 and grade (**Supplementary Methods**): (1) luminal A-like, (2)
547 luminal B/HER2-negative-like, (3) luminal B-like, (4) HER2-enriched-like and (5) TN. We
548 performed sensitivity analyses, in which we estimated the ORs and 95% CI between the variants
549 and the intrinsic-like subtypes by implementing a standard polytomous model restricted to
550 cases with complete tumor marker data. For all analyses we analyzed OncoArray and iCOGS
551 array data separately, adjusting for the first 10 principal components for ancestry-informative
552 variants, and then meta-analyzed the results. To help describe sources of heterogeneity, we
553 performed cluster analyses based on Euclidean distance calculated from z-statistics that were
554 estimated by the individual marker-specific tumor heterogeneity tests and by the case-control

555 tests between variants and the intrinsic-like subtypes. Clustering was performed in R using the

556 function Heatmap as implemented by the package "Complex Heatmap" version 3.1[34].

557

558 **Data Availability**

559

560 Requests for data can be made to the corresponding author or the Data Access Coordination

561 Committees (DACCs) of BCAC (see above URL). BCAC DACC approval is required to access data

562 from the ABCFS, ABCS, ABCS-F, ABCTB, AHS, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UT,

563 BCINIS, BIGGS, BSUCH, CBCS, CECILE, CNIO-BCS, CTS, DIETCOMPLYF, ESTHER, FHRISK, GC-HBOC,

564 GENICA, GEPARSIXTO, GESBC, HABCS, HCSC, HEBCS, HUBCS, KARBAC, KARMA, KBCP,

565 KCONFAB/AOCS, LMBC, MABCS, MARIE, MBCSG, MCBCS, MISS, MMHS, MTLGEBCS, NBCS, NC-

566 BCFR, NCBCS, OBCS, OFBCR, ORIGO, PKARMA, POSH, PREFACE, PROCAS, RBCS, SASBAC, SBCS,

567 SKKDKFZS, SUCCESSB, SUCCESSC, SZBCS, TNBCC, UCIBCS, UKBGS, UKOPS and USRT studies

568 (**Supplementary Table 1**).

569

570

Funding and Acknowledgments

This project has been funded in part with Federal funds from the National Cancer Institute Intramural Research Program, National Institutes of Health. Dr. Nilanjan Chatterjee was supported by NHGRI (1R01 HG010480-01). OncoArray genotyping was funded by the government of Canada through Genome Canada and the Canadian Institutes of Health Research (GPH-129344), the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Génome Québec, the Quebec Breast Cancer Foundation for the PERSPECTIVE project, the US National Institutes of Health (NIH) (1 U19 CA 148065 for the Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract HHSN268201200008I), Cancer Research UK (C1287/A16563), the Odense University Hospital Research Foundation (Denmark), the National R&D Program for Cancer Control–Ministry of Health and Welfare (Republic of Korea) (1420190), the Italian Association for Cancer Research (AIRC; IG16933), the Breast Cancer Research Foundation, the National Health and Medical Research Council (Australia) and German Cancer Aid (110837).

iCOGS genotyping was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710, C1287/A10118 and C12292/A11174]), NIH grants (CA128978, CA116167 and CA176785) and the Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 (GAME-ON initiative)), an NCI Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, the Ministère de

l'Économie, Innovation et Exportation du Québec (PSR-SIIRI-701), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund.

A full description of the funding and acknowledgments is provided in the **Supplementary Information**.

596

597 **References**

598 1. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature*
599 **490**, 61-70 (2012).

600

601 2. Curigliano G*, et al.* De-escalating and escalating treatments for early-stage breast cancer: the St.
602 Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast
603 Cancer 2017. *Ann Oncol* **28**, 1700-1712 (2017).

604

605 3. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic
606 tumor subtypes. *Biochim Biophys Acta* **1856**, 73-85 (2015).

607

608 4. Yang XR*, et al.* Associations of breast cancer risk factors with tumor subtypes: a pooled analysis
609 from the Breast Cancer Association Consortium studies. *Journal of the National Cancer Institute*
610 **103**, 250-263 (2011).

611

612 5. Michailidou K*, et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**,
613 92-94 (2017).

614

615 6. Milne RL*, et al.* Identification of ten variants associated with risk of estrogen-receptor-negative
616 breast cancer. *Nature genetics* **49**, 1767-1778 (2017).

617

618 7. Garcia-Closas M*, et al.* Genome-wide association studies identify four ER negative-specific breast
619 cancer risk loci. *Nature genetics* **45**, 392-398, 398e391-392 (2013).

620

621 8. Dunning AM*, et al.* Breast cancer risk variants at 6q25 display different phenotype associations
622 and regulate ESR1, RMND1 and CCDC170. *Nature genetics* **48**, 374-386 (2016).

623

624 9. Milne RL*, et al.* Confirmation of 5p12 as a susceptibility locus for progesterone-receptor-
625 positive, lower grade breast cancer. *Cancer epidemiology, biomarkers & prevention : a*
626 *publication of the American Association for Cancer Research, cosponsored by the American*
627 *Society of Preventive Oncology* **20**, 2222-2231 (2011).

628

629 10. Figueroa JD*, et al.* Associations of common variants at 1p11.2 and 14q24.1 (RAD51L1) with
630 breast cancer risk and heterogeneity by tumor subtype: findings from the Breast Cancer
631 Association Consortium. *Human molecular genetics* **20**, 4693-4706 (2011).

632

633 11. Orr N*, et al.* Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2.
634 *Human molecular genetics* **24**, 2966-2984 (2015).

635

12. Broeks A*, et al.* Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Human molecular genetics* **20**, 3289-3303 (2011).

13. Zhang H, Zhao N, Ahearn TU, Wheeler W, García-Closas M, Chatterjee N. A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *Biostatistics*, (2020).

14. Zhang H*, et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *bioRxiv*, 778605 (2019).

15. Easton DF*, et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-1093 (2007).

16. Hunter DJ*, et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics* **39**, 870-874 (2007).

17. Stacey SN*, et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature genetics* **39**, 865-869 (2007).

18. Fachal L*, et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature genetics* **52**, 56-73 (2020).

19. Li J*, et al.* Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. *Mol Cell* **9**, 1045-1054 (2002).

20. McKay JD*, et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature genetics* **49**, 1126-1132 (2017).

21. Wang Y*, et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nature genetics* **46**, 736-741 (2014).

22. Bojesen SE*, et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics* **45**, 371-384, 384e371-372 (2013).

23. Li X*, et al.* Association of multiple genetic variants with breast cancer susceptibility in the Han Chinese population. *Oncotarget* **7**, 85483-85491 (2016).

24. Darabi H*, et al.* Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. *American journal of human genetics* **97**, 22-34 (2015).

677
678 25. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological
679 grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*
680 **19**, 403-410 (1991).

681
682 26. Mazoyer S*, et al.* A polymorphic stop codon in BRCA2. *Nature genetics* **14**, 253-254 (1996).

683
684 27. Meeks HD*, et al.* BRCA2 Polymorphic Stop Codon K3326X and the Risk of Breast, Prostate, and
685 Ovarian Cancers. *Journal of the National Cancer Institute* **108**, (2016).

686
687 28. Darabi H*, et al.* Fine scale mapping of the 17q22 breast cancer locus using dense SNPs,
688 genotyped within the Collaborative Oncological Gene-Environment Study (COGs). *Scientific*
689 *reports* **6**, 32512 (2016).

690
691 29. Michailidou K*, et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer
692 risk. *Nature genetics* **45**, 353-361, 361e351-352 (2013).

693
694 30. Little RJA, Rubin DB. Statistical analysis with missing data. In: *Wiley series in probability and*
695 *statistics*). Third edition edn. Wiley, (2019).

696
697 31. Mavaddat N*, et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer
698 Subtypes. *American journal of human genetics* **104**, 21-34 (2019).

699
700 32. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via Em
701 Algorithm. *J Roy Stat Soc B Met* **39**, 1-38 (1977).

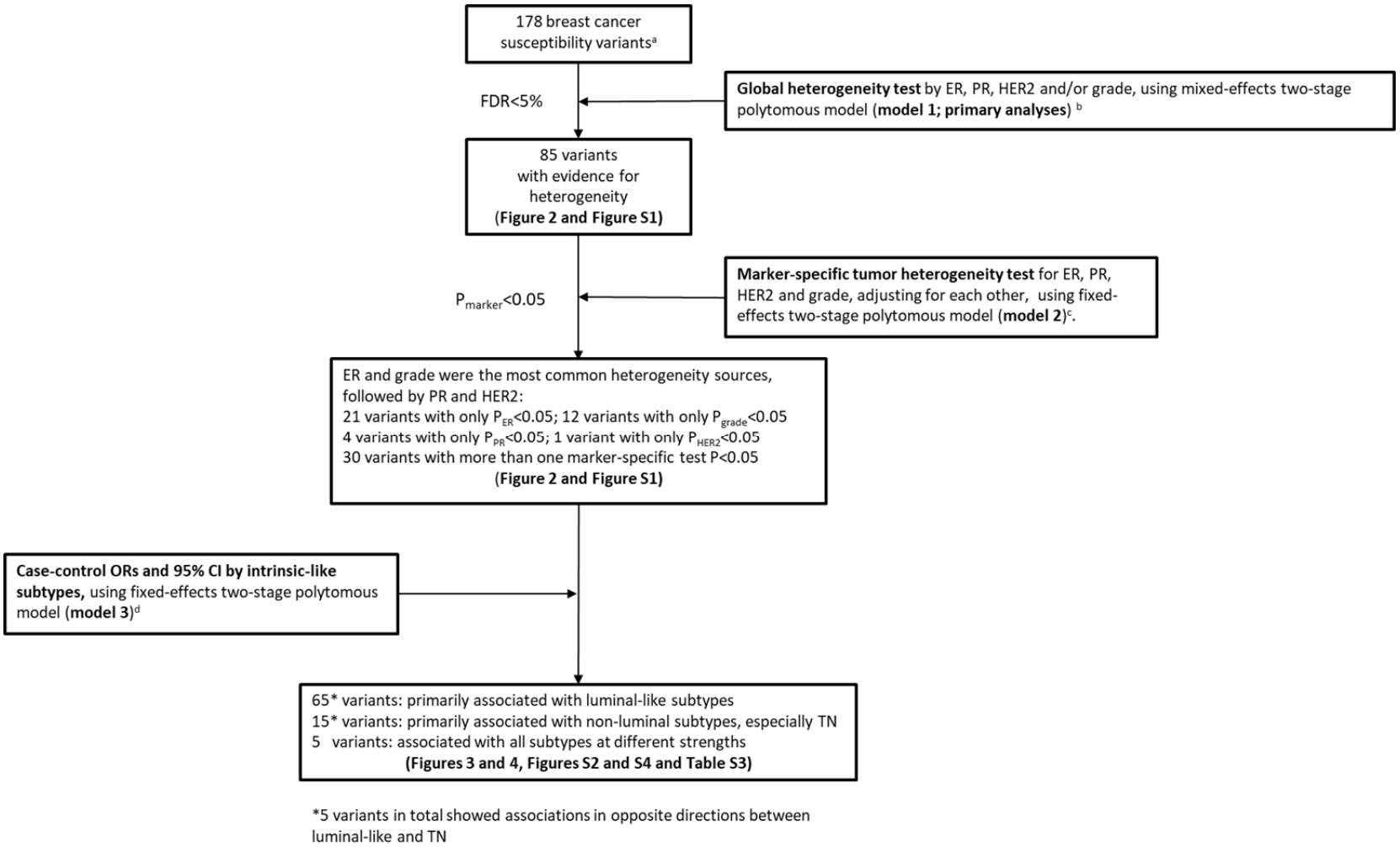
702
703 33. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful
704 Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*
705 **57**, 289-300 (1995).

706
707 34. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in
708 multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

709

710

Figure 1. Overview of the analytic strategy and results from the investigation of 178 known breast cancer susceptibility variants for evidence of heterogeneity according to the estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and grade



a=We evaluated 178 variants reported by Michailidou k, et al Nature 2017 and Milne RL, et al Nature Genetics 2017 that were associated with breast cancer at a p-value of $<10^{-4}$

b=Model 1 (Primary analyses): Mixed-effect two-stage polycotomos model (ER as fixed-effect, and PR, HER2 and grade as random-effects) to test for global heterogeneity between each individual susceptibility variant and any of the tumor features (separate models were fit for each variant).

c=Model 2: Fixed effect two-stage polycotomos model for marker-specific tumor heterogeneity tests between each individual variant and each of the tumor features (ER, PR, HER2, and grade) mutually adjusted for each other (separate models were fit for each variant).

d=Model 3: Fixed effect two-stage polycotomos model for intrinsic-like subtypes: luminal A-like, luminal B/HER2-negative-like, luminal B-like, HER2-enriched-like, and basal-like/triple negative.

Note: In extended analyses, we fitted similar mixed and fixed effects models to evaluate heterogeneity according to ER, PR, HER2, grade and/or triple-negative status (TN; **Supplementary Fig. 3**)

Figure 2. Heatmap and clustering of the P-values from the two-stage model's case-case marker-specific heterogeneity test for associations between each of the 178 breast cancer susceptibility variants and estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) or grade, adjusting for principal components and each tumor marker. Columns represent individual variants. For more detailed information on the context of figure see **Supplementary Fig. 1.**

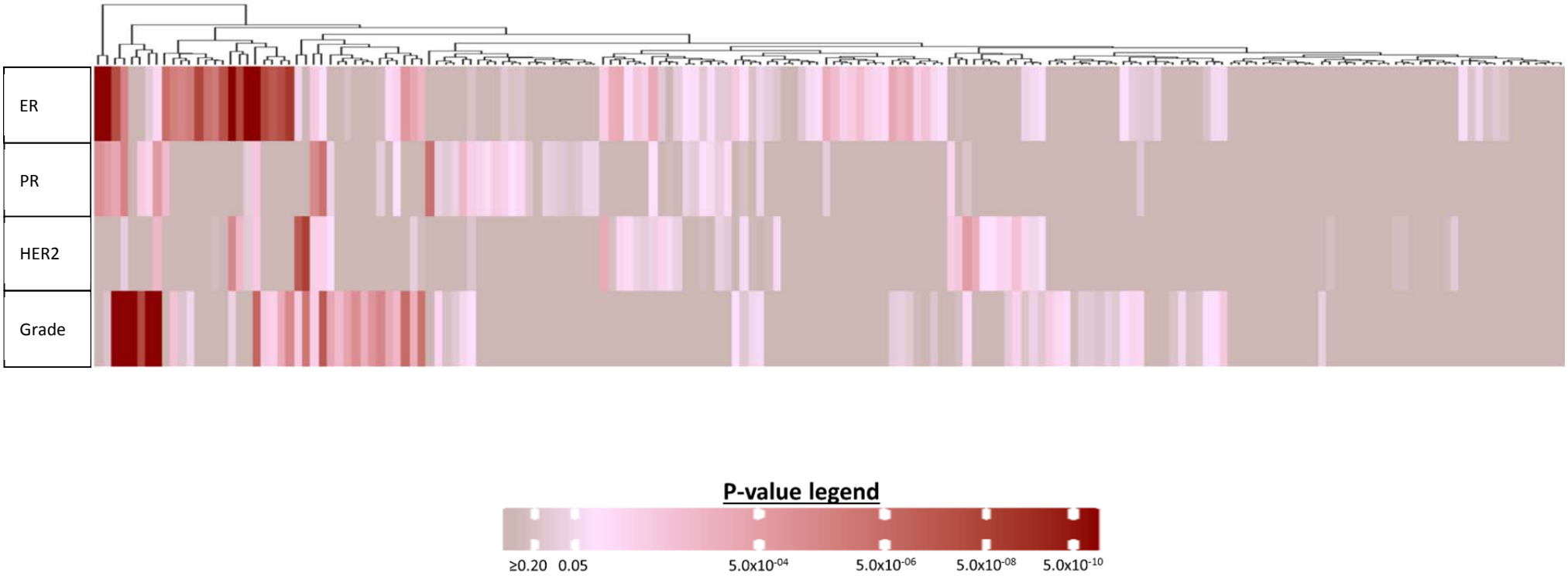
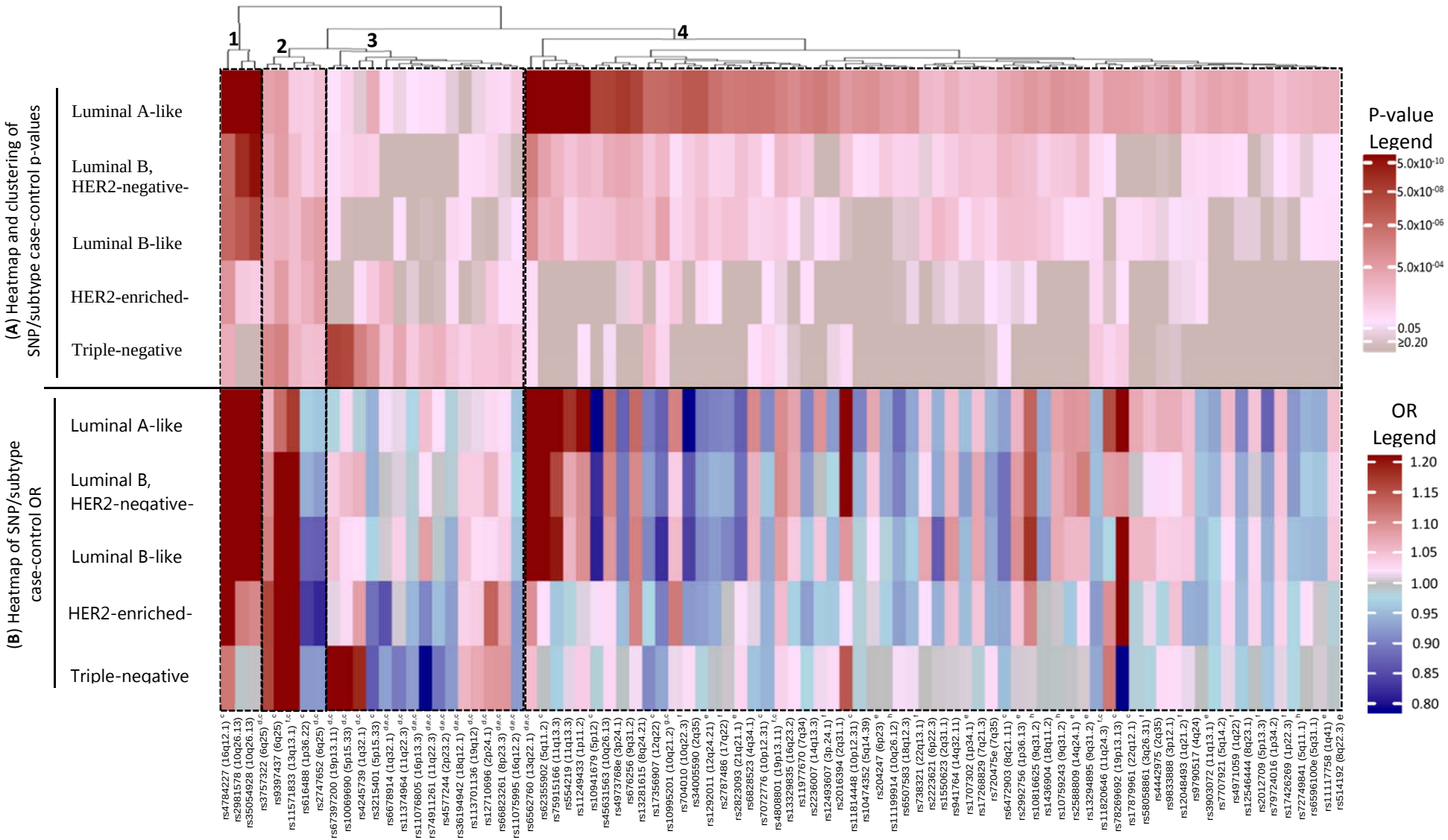


Figure 3. Heatmap and clustering of case-control P-values and odds ratios (OR) from two-stage models for intrinsic-like subtypes, fitted for each of the 85 breast cancer susceptibility variants with evidence for global heterogeneity. Columns are labeled according to variants and loci. (**A**) Clustering based on case-control p-value of the associations between susceptibility variants and breast cancer intrinsic-like subtypes[a,b]. (**B**) Heatmap of ORs between susceptibility variants and



intrinsic-like subtypes. For more detailed information on the context of figure see **Supplementary Fig. 2.**

a = luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); luminal B/HER2-negative-like (ER+ and/or PR+, HER2-, grade 3); luminal B-like (ER+ and/or PR+, HER2+); HER2-enriched-like (ER- and PR-, HER2+); TN (ER-, PR-, HER2-)

b = Clusters: 1 and 4: strongest associations with hormone receptor (HR)-positive subtypes; Cluster 2: Associated with all subtypes at varying strengths, and Cluster 3: strongest associations with TN or non-luminal subtypes

c = 32 variants associated ($p<0.05$) with risk of the TN subtype

d = variants previously identified as primarily predisposing to ER-negative or TN disease as reported in Milne RL, et al. Nat Genet 2017;49(12):1767-1778

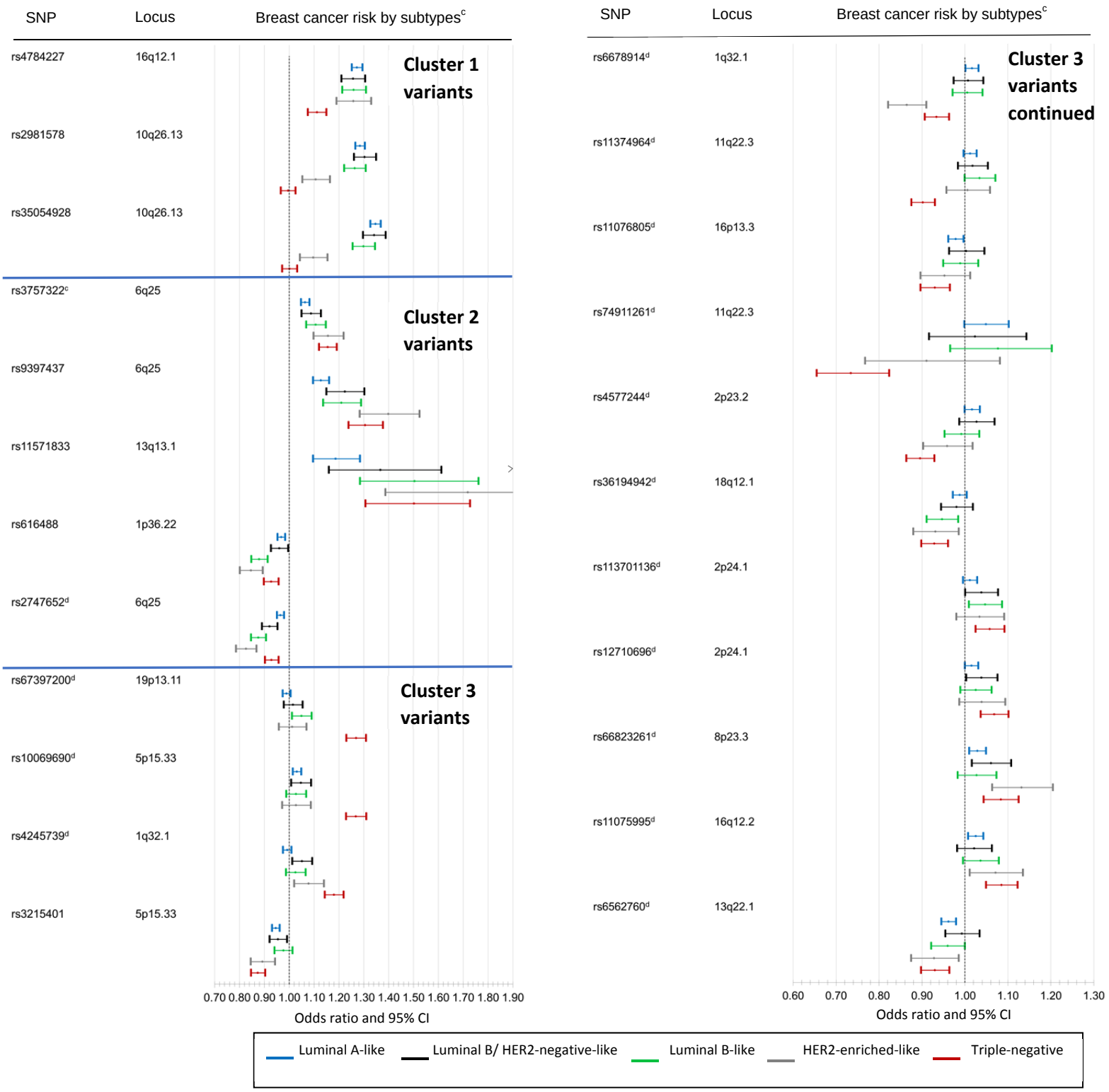
e = 21 variants with evidence for having ER as the only source of heterogeneity

f = 12 variants with evidence for having grade as the only source of heterogeneity

g = 1 variants with evidence for having HER2 as the only source of heterogeneity

h = 4 variants with evidence for having PR as the only source of heterogeneity

Figure 4. Associations[a] between breast cancer susceptibility variants with evidence of heterogeneity in the two-stage model[b] and intrinsic-like subtypes. variants presented in order as shown in **figure 3**, clusters 1, 2 and 3.



[a] = Per-minor allele odds ratio (95% confidence limits).

[b] = Two-stage model testing for heterogeneity according to estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and grade

[c] = luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); luminal B/HER2-negative-like (ER+ and/or PR+, HER2-, grade 3); luminal B-like (ER+ and/or PR+, HER2+); (4) HER2-enriched-like (ER- and PR-, HER2+), and triple-negative (ER-, PR-, HER2-)

[d] = variants previously identified as primarily predisposing to ER-negative or TN disease as reported in Milne RL, et al. Nat Genet 2017;49(12):1767-1778.

Table 1. Distribution of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and grade and the intrinsic-like subtypes [a,b] among invasive cases of breast cancer in studies from the Breast Cancer Consortium Association.

Tumor Marker	N (%)
ER	
Negative	16,900 (19)
Positive	70,030 (81)
Unknown	19,641
PR	
Negative	24,283 (32)
Positive	51,603 (68)
Unknown	30,685
HER2	
Negative	47,693 (83)
Positive	9,529 (17)
Unknown	49,349
Grade	
1	15,583 (20)
2	37,568 (49)
3	24,382 (31)
Unknown	29,038
Intrinsic-like subtypes	
Luminal A-like	33,083 (59)
Luminal B/Her2-negative-like	6,804 (12)
Luminal B-like	6,511 (12)
HER2-enriched-like	2,797 (5)
Triple-negative	7,178 (13)

a=luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); luminal B/HER2-negative-like (ER+ and/or PR+, HER2-, grade 3); luminal B-like (ER+ and/or PR+, HER2+); HER2-enriched-like (ER- and PR-, HER2+); TN (ER-, PR-, HER2-)
b=Intrinsic subtypes defined among 56,373 cases with available tumor marker data