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 Hamann[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.

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

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

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

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

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

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

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

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

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

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

*These authors contributed equally to this work

**These authors contributed equally to this work

†Lists of participants and their affiliations appear in the Funding and Acknowledgments

Financial support: see Supplementary Information

Conflicts of interest: The authors declare no potential conflicts of interest.

Key Words: Breast Cancer, etiologic heterogeneity, genetic predisposition, common breast cancer susceptibility variants

Corresponding Author

Montserrat Garcia-Closas

National Cancer Institute

Division of Cancer Epidemiology & Genetics

9609 Medical Center Drive

Rockville, MD 20850

montserrat.garcia-closas@nih.gov

Abstract

Genome-wide association studies (GWAS) have identified multiple common breast cancer susceptibility variants. 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. Among 106,571 invasive breast cancer cases and 95,762 controls of European ancestry with data on 173 breast cancer variants identified in previous GWAS, we used novel 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 173 variants were associated with at least one tumor feature (false discovery rate <5%), most commonly ER and grade, followed by PR and HER2. Models for intrinsic-like subtypes found nearly all of these variants (83 of 85) associated at P<0.05 with risk for at least one luminal-like subtype, and approximately half (41 of 85) of the variants were associated with risk of at least one non-luminal subtype, including 32 variants associated with triple-negative (TN) disease. Ten variants were associated with risk of all subtypes in different magnitude. 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 variants and can inform investigations of subtype-specific risk prediction.

Significance

Our results demonstrate complex etiologic heterogeneity patterns of common breast cancer susceptibility variants and can inform functional analyses to identify underlying causal variants and the development of subtype-specific polygenic risk scores.

Introduction

Breast cancer represents a heterogenous group of diseases with different molecular and clinical features(1). Clinical assessment of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and histological grade are routinely determined to inform treatment strategies and prognostication(2). Combined, these tumor features define five intrinsic-like subtypes (i.e., luminal A-like, luminal B–like/HER2-negative, luminal B-like/HER2-positive, HER2-positive/non-luminal, and triple negative) that are correlated with intrinsic subtypes defined by gene expression panels(2,3) Most known breast cancer risk or protective factors are related to luminal or hormone receptor (ER or PR) positive tumors, whereas less is known about the etiology of triple-negative (TN) tumors, an aggressive subtype(4,5).

Breast cancer genome-wide association studies (GWAS) have identified over 170 common susceptibility variants, most of them single nucleotide polymorphisms (SNPs), of which many are differentially associated with ER-positive than ER-negative disease(6-8). These include 20 variants that primarily predispose to ER-negative or TN disease(7,8). However, few studies have evaluated variant associations with other tumor features, or simultaneously studied multiple, correlated tumor markers to identify source(s) of etiologic heterogeneity(7,9-13). We recently developed a two-stage polytomous logistic regression method that efficiently characterizes etiologic heterogeneity while accounting for tumor marker correlations and missing tumor data(14,15). This method can help describe complex relationships between susceptibility variants and multiple tumor features, helping to clarify breast cancer subtype 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(15).

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 with 173 breast cancer risk variants identified by GWAS(6,7) by tumor subtypes defined by ER, PR, HER2 and tumor grade.

Methods

Study Population and Genotyping

The study population and genotyping are described in previous publications(6,7) and in the **Supplementary Methods**. We included invasive cases and controls from 81 BCAC studies with genotyping data from two Illumina genome-wide custom arrays, the iCOGS and OncoArray (106,571 cases (OncoArray: 71,788; iCOGS: 34,783) and 95,762 controls (OncoArray: 58,134; iCOGS: 37,628); **Supplementary Table 1**). We evaluated 173 breast cancer risk variants that were identified in or replicated by prior BCAC analyses to be associated with breast cancer risk at a p-value threshold $p<5.0x<10^{-8}$ (6,7). Most of these variants (n=153) were identified because of their association with risk of overall breast cancer, and a small number of variants (n=20) were identified because of their association specific to ER-negative breast cancer (**Supplementary Table 2**). These 173 variants have not previously been simultaneously investigated for evidence of tumor heterogeneity with multiple tumor

markers(6,7,15,16). Genotypes for the variants marking the 173 susceptibility loci were determined by genotyping with the iCOGS and the OncoArray arrays and imputation to the 1000 Genomes Project (Phase 3) reference panel.

Statistical Analysis

An overview of the analytic strategy is shown in **Figure** 1 and a detailed discussion of the statistical methods, including the two-stage polytomous logistic regression, are provided in the **Supplementary Methods** and elsewhere(14,15). Briefly, we used two-stage polytomous regression models that allow modelling of genetic association of breast cancer accounting for underlying heterogeneity in associations by combinations of multiple tumor markers using a parsimonious decomposition of subtype-specific case-control odds-ratio parameters in terms of marker-specific case-case odd-ratio parameters(14,15). We introduced further parsimony by using mixed-effect formulation of the model that allows ER-specific case-case parameters to be treated as fixed and similar parameters for other markers (PR, HER2 and grade) as random. We used an expectation–maximization (EM) algorithm(17) for parameter estimation under this model to account for missing data in tumor characteristics. A detailed description of the two-stage polytomous regression models used in this manuscript is presented in the **Supplementary Methods** and in separate manuscripts(14,15).

Our primary aim was to identify which of 173 known breast cancer susceptibility variants showed heterogenous risk associations by ER, PR, HER2 and grade. This was tested using a global heterogeneity test by ER, PR, HER2 and/or grade, with a mixed-effect two-stage polytomous model (**model 1**), fitted separately for each variant. The global null hypothesis was

that there was no difference in risk of breast cancer associated with the variant genotype across any of the tumor features being evaluated. We accounted for multiple testing (173 tests, one for each of variant) of the global heterogeneity test using a false discovery rate (FDR) <0.05 under the Benjamini-Hochberg procedure(18).

For the variants showing evidence of global heterogeneity after FDR adjustment, we further evaluated which of the tumor features contributed to the heterogeneity by fitting a fixed-effects two-stage model (**model 2**) that simultaneously tested for associations with each tumor feature (this model was fitted for each variant separately). We used a threshold of P<0.05 for marker-specific tumor heterogeneity tests to describe which specific tumor marker(s) contributed to the observed heterogeneity, adjusting for the other tumor markers in the model. This p-value threshold was used only for descriptive purposes, as the primary hypotheses were tested using the FDR-adjusted global test for heterogeneity described above.

We conducted additional analyses to explore evidence of heterogeneity. We fitted a fixed-effect two-stage model (**model 3**) to estimate case-control odd ratios (ORs) and 95% confidence intervals (CI) between the variants and five intrinsic-like subtypes defined by combinations of ER, PR, HER2 and grade: (1) luminal A-like (ER+ and/or PR+, HER2-, grade 1 or 2); (2) luminal B-like/HER2-negative (ER+ and/or PR+, HER2-, grade 3); (3) luminal B-like/HER2-positive (ER+ and/or PR+, HER2+); (4) HER2-positive/non-luminal (ER- and PR-, HER2+), and (5) TN (ER-, PR-, HER2-). We also fitted a fixed-effect two-stage model to estimate case-control ORs and 95% confidence intervals (CI) with tumor grade (**model 4**; defined as grade 1, grade 2, and grade 3) for the variants associated at P<0.05 only with grade in case-case comparisons from model 2.

To help describe sources of heterogeneity from different tumor characteristics in models 2 and 3, we performed cluster analyses based on Euclidean distance calculated from z-statistics that were estimated by the individual marker-specific tumor heterogeneity tests (model 2) and the case-control associations with risk of intrinsic-like subtypes (model 3) . The clusters were used only to help present our findings and were not intended to suggest strictly defined categories. Clustering was performed in R using the function Heatmap as implemented by the package "Complex Heatmap" version 3.1(19).

We also tested for evidence of heterogenous risk associations by TN status by fitting an extended mixed-effect two-stage polytomous model (model 1-extended), using a global heterogeneity test by ER, PR, HER2, TN, and/or grade, and an extended fixed-effects two stage model polytomous model to test which tumor features, ER, PR, HER2, TN, and/or grade, contributed to the heterogeneity (model 2-extended).

We performed sensitivity analyses, in which we estimated the ORs and 95% CI between the variants and the intrinsic-like subtypes by implementing a standard polytomous model restricted to cases with complete tumor marker data. For all analyses we analyzed OncoArray and iCOGS array data separately, adjusting for the first 10 principal components for ancestry-informative variants, and then meta-analyzed the results.

Results

The mean (SD) ages at diagnosis (cases) and enrollment (controls) were 56.6 (12.2) and 56.4 (12.2) years, respectively. Among cases with information on the corresponding tumor marker, 81% were ER-positive, 68% PR-positive, 83% HER2-negative and 69% grade 1 or 2 (**Table 1;** see **Supplementary Table 1** for details by study). **Supplementary Table 3** shows the

correlation between the tumor markers. ER was positively correlated with PR (r^2=0.61) and inversely correlated with HER2 (r^2=-0.16) and grade (r^2=-0.39). The most common intrinsic-like subtype was luminal A-like (54%), followed by TN (14%), luminal B-like/HER2-negative (13%), Luminal B-like/HER2-positive (13%) and HER2-positive/non-luminal (6%; **Table 1**).

Figure 1 shows an overview of the analytic strategy and results from three main analyses performed separately for each variant: 1) global test for heterogeneity by all tumor markers (model 1; primary hypothesis), 2) marker-specific tumor test for heterogeneity for each marker, adjusting for the others (model 2), 3) estimation of case-control ORs (95%CIs) by intrinsic-like subtypes (model 3), and 4) estimation of case-control ORs (95%CIs) by tumor grade (model 4).

1) Global test for heterogeneity by tumor markers (primary hypothesis)

Mixed-effects two-stage models (model 1) were fitted for each of the 173 variants separately and included terms for ER, PR, HER2 and grade to test of global heterogeneity by any of the tumor features (case-case comparison). This model identified 85 of 173 (49.1%) variants with evidence of heterogeneity by at least one tumor feature (FDR<5%; **Figure 1-2; Supplementary Fig. 1**).

2) Marker-specific tumor test for heterogeneity for each marker, adjusting for other markers

Fixed-effects two-stage models (model 2) was used to test which of the correlated tumor markers was responsible for the observed global heterogeneity in the 85 variants (case-

case comparison). These analyses identified ER and grade as the two features that most often contributed to the observed heterogeneity (45 and 33 variants had marker-specific P<0.05 for ER and grade, respectively), and 29 variants were associated with more than one tumor feature (**Figure 1**, **Supplementary Fig. 1)**. Eighteen of these 85 variants showed no associations with any individual tumor marker at P<0.05 (**Supplementary Fig. 1**). Twenty-one variants were associated at P<0.05 only with ER, 12 variants only with grade, 4 variants only with PR and one variant only with HER2 (**Supplementary Fig. 1,** see footnotes).

3) Estimation of case-control ORs (95%CIs) by intrinsic-like subtypes (model 3)

Fixed-effects two-stage models for intrinsic-like subtypes (model 3) were fitted for each of the 85 variants with evidence of global heterogeneity to estimate ORs (95% CIs) for risk associations with each subtype (case-control comparisons). **Supplementary Figure 2** shows a summary of these analyses for the 85 variants, clustered by case-control p-value of association between susceptibility variants and breast cancer intrinsic-like subtypes, and **Supplementary Figures 3** shows forest plots for associations with risk by tumor subtypes. Nearly all (83 of 85) variants were associated with risk (P<0.05) for at least one luminal-like subtype, and approximately half (41 of 85) of the variants were associated with risk of at least one non-luminal subtype, including 32 variants that were associated with TN disease (**Figure 1**, **Supplementary Fig. 2 footnote 'h'**). Ten variants were associated with risk of all subtypes (**Figure 1**, **Supplementary Fig. 2 footnote 'j'**). Below we describe examples of groups of variants associated with different patterns of associations with intrinsic subtypes (**Figure 3 a-d**).

Two correlated (r^2=0.73) variants at 10q26.13 (rs2981578 and rs35054928) and

16q12.1-rs4784227 had the strongest evidence of association with risk of luminal-like subtypes (**Figure 3a**, **Supplementary Fig. 2).** The two variants at 10q26.13 showed no evidence of associations with TN subtypes, and a weaker association with HER2-positive/non-luminal subtype (**Figure 3a**, **Supplementary Fig. 2**). In an extended two-stage model (model 2-extended) to specifically test for heterogeneity between TN vs non-TN subtypes, both rs2981578 and rs35054928 were strongly associated with TN status (**Supplementary Fig. 4**). In contrast, 16q12.1-rs4784227 was strongly associated with risk for all luminal-like subtypes and, weaker so, with risk of HER2-positive/non-luminal and TN subtypes (**Figures 3a**, **Supplementary Fig. 2**).

Three variants 19p13.11-rs67397200, 5p15.33-rs10069690 and 1q32.11-rs4245739 showed the strongest evidence of associations with risk of TN disease (**Figure 3b**, **Supplementary Fig. 2**). In the model 2-extended these variants were significantly associated with TN status (**Supplementary Fig. 4**).

Two weakly correlated variants in 6q25 (r^2=0.17), rs9397437 and rs3757322, and a third variant in 6q25, rs2747652, which was not correlated (r^2<0.01) with rs9397437 or rs3757322, showed strong evidence of being associated with risk of all subtypes. rs9397437 and rs3757322 had strong evidence of associations with risk of TN and risk of luminal-like subtypes (**Figures 3c** and **Supplementary Fig. 2**). rs2747652 was most strongly associated with risk of the HER2-positive/non-luminal subtype (**Figures 3c**, **Supplementary Fig. 2**).

Five variants were associated with risk of luminal A-like disease in an opposite direction to their association with risk of TN disease (**Figure 3d**, **Supplementary Fig. 2**). 1q32.1-rs6678914, 2p23.2-rs4577244, and 19p13.11-rs67397200 had weaker evidence of associations

with risk of luminal A-like disease compared to associations with risk of TN disease, and 10p12.31-rs7072776 and 22q12.1-rs17879961 (I157T) had stronger evidence of an association with risk of luminal A-like disease compared to their association with risk of TN disease (**Figure 3d**, **Supplementary Fig. 2**, for rs67397200 see **Figure 3b**). In model 2-extended, rs7072776, rs17879961, and rs67397200 were significantly associated with TN status (**Supplementary Fig. 4**).

4) Estimation of case-control ORs (95%CIs) by tumor grade (model 4)

Case-control associations by tumor grade for the 12 variants associated at P<0.05 only with grade in case-case comparisons are shown in **Supplementary Fig. 5**. 13q13.1-rs11571833, 1p22.3-rs17426269 and 11q24.3-rs11820646 showed stronger evidence for predisposing to risk of high-grade subtypes, and the remaining variants showed stronger evidence for predisposing to risk of low-grade subtypes.

When limiting analyses to cases with complete tumor marker data, results from case-control analyses were similar, but less precise than results from the two-stage polytomous regression model using the EM algorithm to account for missing tumor marker data (**Supplementary Table 4**).

Discussion

This study demonstrates the extent and complexity of genetic etiologic heterogeneity among 173 breast cancer risk variants by multiple tumor characteristics, using novel

methodology in the largest and the most comprehensive investigation conducted to date. We found compelling evidence that about half of the investigated breast cancer susceptibility loci (85 of 173 variants) predispose to tumors with different characteristics. We identified tumor grade, along with confirming ER and TN status, as important determinants of etiologic heterogeneity. Associations with individual tumor features translated into differential associations with the risk of intrinsic-like subtypes defined by their combinations.

Many of the variants with evidence of global heterogeneity predisposed to risk of multiple subtypes, but with different magnitudes. For example, three variants identified in early GWAS for overall breast cancer, *FGFR2* (rs35054928 and rs2981578)(20,21) and 8q24.21 (rs13281615)(20), were associated with luminal-like and HER2-positive/non-luminal subtypes, but not with TN disease. rs4784227 located near *TOX3(20,22)* and rs62355902 located in a *MAP3K1(20)* regulatory element, were associated with risk of all five subtypes. Of the five variants found associated in opposite directions with luminal A-like and TN disease, we previously reported rs6678914 and rs4577244 to have opposite effects between ER-negative and ER-positive tumors(7). rs17879961 (I157T), a likely causal(16) mis-sense variant located in a *CHEK2* functional domain that reduces or abolishes substrate binding(23), was previously reported to have opposite directions of effects on lung adenocarcinoma and lung squamous cell carcinoma and for lung cancer between smokers and non-smokers(24,25). However, further studies are required to follow-up and clarify the mechanisms for these apparent cross-over effects.

In prior ER-negative GWAS, we identified 20 variants that predispose to ER-negative disease, of which five variants were only or most strongly associated with risk of TN disease

(rs4245739, rs10069690, rs74911261, rs11374964, and rs67397200)(7,8). We confirmed these five variants to be most strongly associated with TN disease. The remaining previously identified 15 variants all showed associations with risk of non-luminal subtypes, especially TN disease, and for all but four variants (rs17350191, rs200648189, rs6569648, and rs322144) evidence of global heterogeneity was observed.

Little is known regarding PR and HER2 as sources of etiologic heterogeneity independent of ER or TN status. Of the four variants that showed evidence of heterogeneity only according to PR, rs10759243(6,26), rs11199914(6) and rs72749841(6) were previously found primarily associated with risk of ER-positive disease, and rs10816625 was found to be associated with risk of ER-positive/PR-positive tumors, but not other ER/PR combinations(12). rs10995201 was the only variant found in case-case comparisons to be solely associated with HER2 status, although the evidence was not strong, requiring further confirmation. Previously, rs10995201 showed no evidence of being associated with ER status(27). Most variants associated with PR or HER2, had not been investigated for PR or HER2 heterogeneity while adjusting for ER(9-13). We previously reported rs10941679 to be associated with PR-status, independent of ER, and also with grade(10). We also found suggestive evidence of PR-specific heterogeneity for 16q12-rs3803662(13), which is in high LD (r^2= 0.78) with rs4784227 (*TOX3*), a variant strongly associated with PR status. Our findings for rs2747652 are also consistent with a prior BCAC fine-mapping analysis across the *ESR1* locus, which found rs2747652 to be associated with risk of the HER2-positive/non-luminal subtype and high grade independent of ER(9). rs2747652 overlaps an enhancer region and is associated with reduced *ESR1* and *CCDC170* expression(9).

Histologic grade is a composite of multiple tumor characteristics including mitotic count, nuclear pleomorphism, and degree of tubule or gland formation(28). Among the 12 variants identified with evidence of heterogeneity by grade only, rs17426269, rs11820646, and rs11571833 were found to be most strongly associated with risk of grade 3 disease. rs11571833 lies in the *BRCA2* coding region and produces a truncated form of the protein(29) and has been shown to be associated with both risk of TN disease and risk of serous ovarian tumors, both of which tend to be high-grade(30). To our knowledge, rs17426269 and rs11820646 have not been investigated in relation to grade heterogeneity. The remaining 9 variants were all more strongly associated with grade 1 or grade 2 disease. Five of these variants were previously reported to be associated primarily with ER-positive disease(6,31,32), highlighting the importance of accounting for multiple tumor characteristics to better illuminate heterogeneity sources.

We identified 18 variants with evidence of global heterogeneity (FDR<5%), but no significant (marker-specific P<0.05) associations with any of the individual tumor characteristic(s). This is likely explained by the fact that the test for association with specific tumor markers using fixed-effects models are less powerful than mixed-effects models used to test the primary hypothesis of global heterogeneity by any tumor marker(14).

To help describe and visualize the strength of the evidence for common heterogeneity patterns, we performed clustered analyses of p-values for tumor marker-specific heterogeneity tests and case-control associations with risk of intrinsic-like subtypes. Because they are based on p-values, these clusters reflect differences in sample size and statistical power to detect associations between variants and specific tumor subtypes. Thus, clusters should not be interpreted as strictly defined categories.

A major strength of our study is our large sample size of over 100,000 breast cancer cases with tumor marker information, and a similar number of controls, making this the largest, most comprehensive breast cancer heterogeneity investigation. Our application of the two-stage polytomous logistic regression enabled adjusting for multiple, correlated tumor markers and accounting for missing tumor marker data. This is a more powerful and efficient modeling strategy for identifying heterogeneity sources among highly correlated tumor markers, compared with standard polytomous logistic regression(14,15). In simulated and real data analyses, we have demonstrated that in the presence of heterogenous associations across subtypes, the two-stage model is more powerful than polytomous logistic regression for detecting heterogeneity. Moreover, we have demonstrated that in the presence of correlated markers, the two-stage model, incorporating all markers simultaneously, has much better ability to distinguish the true source(s) of heterogeneity compared to testing for heterogeneity by analysis of one marker at a time(14,15). In prior analyses, we showed that the two-stage polytomous regression is a powerful approach to identify susceptibility variants that display tumor heterogeneity(15). Notably, in this prior investigation we excluded the genomic regions in which the 173 variants that were investigated in this work are located(15).

Our study also has some limitations. First, many breast cancer cases from studies included in this report had missing information on one or more tumor characteristics. To address this limitation, we implemented an EM algorithm that allowed a powerful analysis to incorporate cases with missing tumor characteristics under the assumption that tumor characteristics are *missing at random* (MAR), i.e., the underlying reason for missing data may depend on observed tumor markers or/and covariate values, but not on the missing values

themselves(33). If this assumption is violated it can lead to an inflated type-one error(14). However, in the context of genetic association testing, the missingness mechanism would also need to be related to the genetic variants under study, which is unlikely. Our study focused on investigating ER, PR, HER2, and grade as heterogeneity sources, and future studies with more detailed tumor characterization could reveal additional etiologic heterogeneity sources.

In summary, our findings provide insights into the complex etiologic heterogeneity patterns of common breast cancer susceptibility loci. These findings may inform future studies, such as fine-mapping and functional analyses to identify the underlying causal variants, clarifying biological mechanisms that drive genetic predisposition to breast cancer subtypes. Moreover, these analyses provide precise estimates of relative risk for different intrinsic-like subtypes that could improve the discriminatory accuracy of subtype-specific polygenic risk scores(34).

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**.

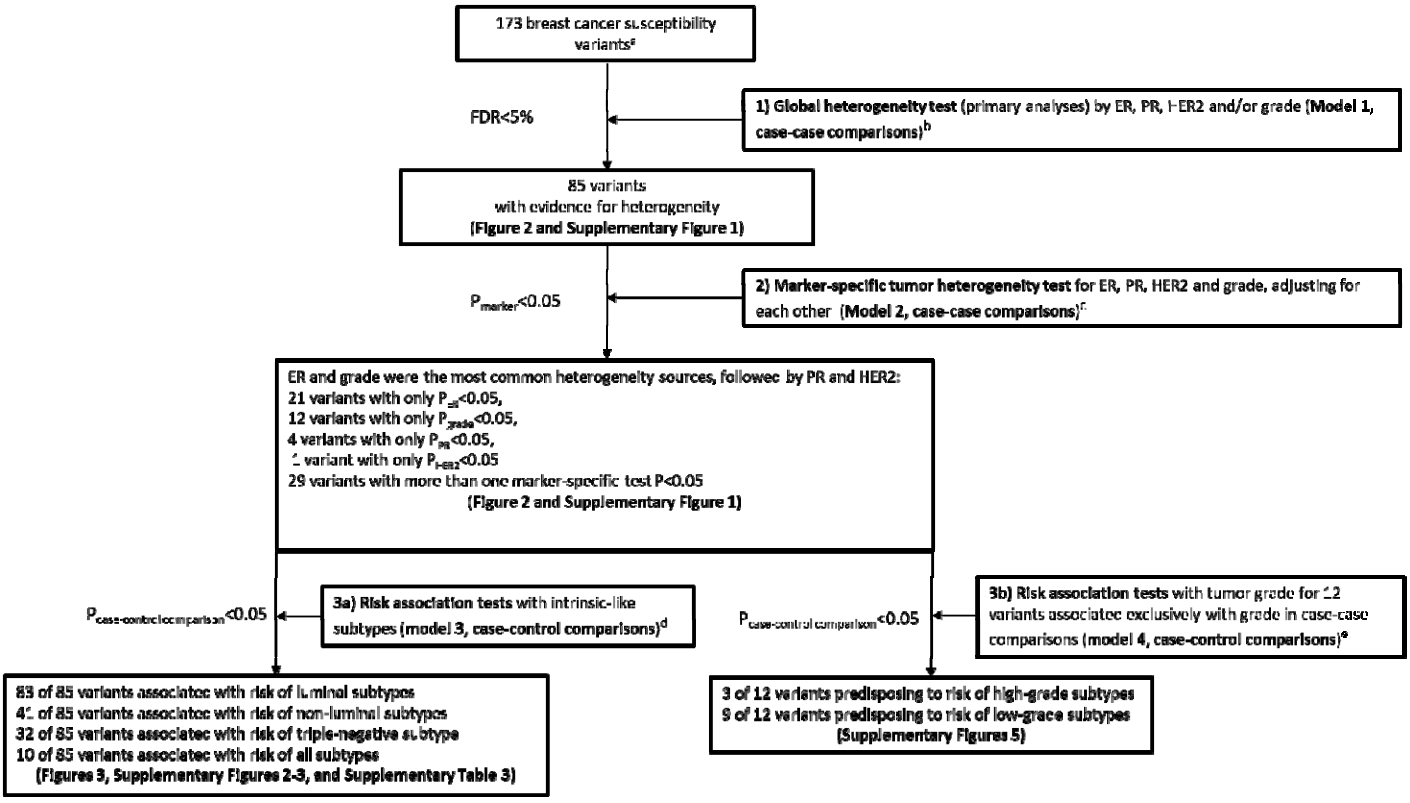
References

1. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature **2012**;490:61-70
2. Curigliano G, Burstein HJ, E PW, Gnant M, Dubsky P, Loibl S, *et al.* De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol **2017**;28:1700-12
3. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, *et al.* Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol **2013**;24:2206-23
4. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. Biochim Biophys Acta **2015**;1856:73-85
5. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, *et al.* Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. Journal of the National Cancer Institute **2011**;103:250-63
6. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, *et al.* Association analysis identifies 65 new breast cancer risk loci. Nature **2017**;551:92-4
7. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nature genetics **2017**;49:1767-78
8. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nature genetics **2013**;45:392-8, 8e1-2
9. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J, *et al.* Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nature genetics **2016**;48:374-86
10. Milne RL, Goode EL, Garcia-Closas M, Couch FJ, Severi G, Hein R, *et al.* Confirmation of 5p12 as a susceptibility locus for progesterone-receptor-positive, lower grade breast cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology **2011**;20:2222-31
11. Figueroa JD, Garcia-Closas M, Humphreys M, Platte R, Hopper JL, Southey MC, *et al.* Associations of common variants at 1p11.2 and 14q24.1 (RAD51L1) with breast cancer risk and heterogeneity by tumor subtype: findings from the Breast Cancer Association Consortium. Human molecular genetics **2011**;20:4693-706
12. Orr N, Dudbridge F, Dryden N, Maguire S, Novo D, Perrakis E, *et al.* Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Human molecular genetics **2015**;24:2966-84
13. Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, Dite GS, *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 **2011**;20:3289-303
14. 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**

15. Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nature genetics **2020**;52:572-81
16. Fachal L, Aschard H, Beesley J, Barnes DR, Allen J, Kar S, *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nature genetics **2020**;52:56-73
17. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via Em Algorithm. J Roy Stat Soc B Met **1977**;39:1-38
18. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) **1995**;57:289-300
19. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics **2016**;32:2847-9
20. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. Nature **2007**;447:1087-93
21. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics **2007**;39:870-4
22. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics **2007**;39:865-9
23. Li J, Williams BL, Haire LF, Goldberg M, Wilker E, Durocher D, *et al.* Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. Mol Cell **2002**;9:1045-54
24. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature genetics **2017**;49:1126-32
25. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nature genetics **2014**;46:736-41
26. Li X, Zou W, Liu M, Cao W, Jiang Y, An G, *et al.* Association of multiple genetic variants with breast cancer susceptibility in the Han Chinese population. Oncotarget **2016**;7:85483-91
27. Darabi H, McCue K, Beesley J, Michailidou K, Nord S, Kar S, *et al.* Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. American journal of human genetics **2015**;97:22-34
28. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology **1991**;19:403-10
29. Mazoyer S, Dunning AM, Serova O, Dearden J, Puget N, Healey CS, *et al.* A polymorphic stop codon in BRCA2. Nature genetics **1996**;14:253-4
30. Meeks HD, Song H, Michailidou K, Bolla MK, Dennis J, Wang Q, *et al.* BRCA2 Polymorphic Stop Codon K3326X and the Risk of Breast, Prostate, and Ovarian Cancers. Journal of the National Cancer Institute **2016**;108
31. Darabi H, Beesley J, Droit A, Kar S, Nord S, Moradi Marjaneh M, *et al.* Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Scientific reports **2016**;6:32512
32. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics **2013**;45:353-61, 61e1-2

33. Little RJA, Rubin DB. Statistical analysis with missing data. Wiley series in probability and statistics. Third edition ed. Hoboken, NJ: Wiley,; 2019. p 1 online resource.
34. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A*, et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. American journal of human genetics **2019**;104:21-34

Figure 1. Overview of the analytic strategy and results from the investigation of 173 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 173 breast cancer risk variants identified in or replicated by prior BCAC GWAS (6,7), see **Methods** and **Supplementary Methods** sections for more details.

[b] Model 1 (primary analyses): Mixed-effect two-stage polytomous model (ER as fixed-effect, and PR, HER2 and grade as random-effects) for global heterogeneity tests (i.e. case-case comparisons from stage 2 of the two-stage model) between each individual risk variant and any of the tumor features (separate models were fit for each variant). In an extended analysis we fit, Model 1-extended, a mixed-effects two-stage polytomous model to test for global heterogeneity between each individual susceptibility variant and ER and triple-negative status (as fixed effects) and PR, HER2, and grade (as random effects).

[c] Model 2: Fixed-effect two-stage polytomous model for marker-specific tumor heterogeneity tests (i.e. case-case comparisons from stage 2 of the two-stage model) 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). In an extended analysis we fit Model 2-extended, a fixed-effects two-stage polytomous model to test for marker-specific tumor heterogeneity between each individual susceptibility variant and each of ER, PR, HER2, grade, and triple-negative status, mutually adjusting for each other.

[d] Model 3: Fixed effect two-stage polytomous model for risk associations with intrinsic-like subtypes (i.e. case-control comparisons from stage 1 of the two-stage model): luminal A-like, luminal B-like/HER2-negative, luminal B-like/HER2-positive, HER2-positive/non-luminal, and triple negative.

[e] Model 4: Fixed effect two-stage polytomous model for risk associations with tumor grade (i.e. case-control comparisons from stage 1 of the two-stage model) for the 12 variants associated at P<0.05 only with grade in case-case comparisons (from model 2): grade 1, grade 2, and grade 3.

Figure 2. Heatmap of the P-values from the fixed-effects two-stage polytomous model for marker-specific heterogeneity tests (case-case comparison from model 2) for association between each of the 173 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.**

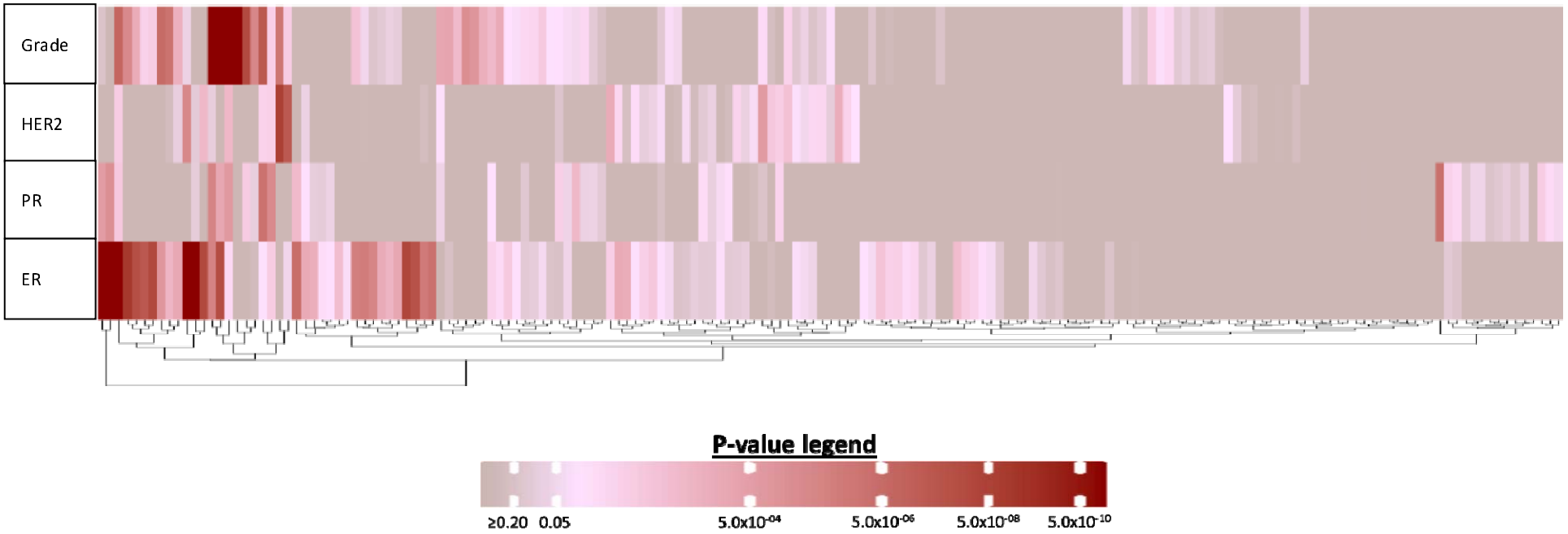
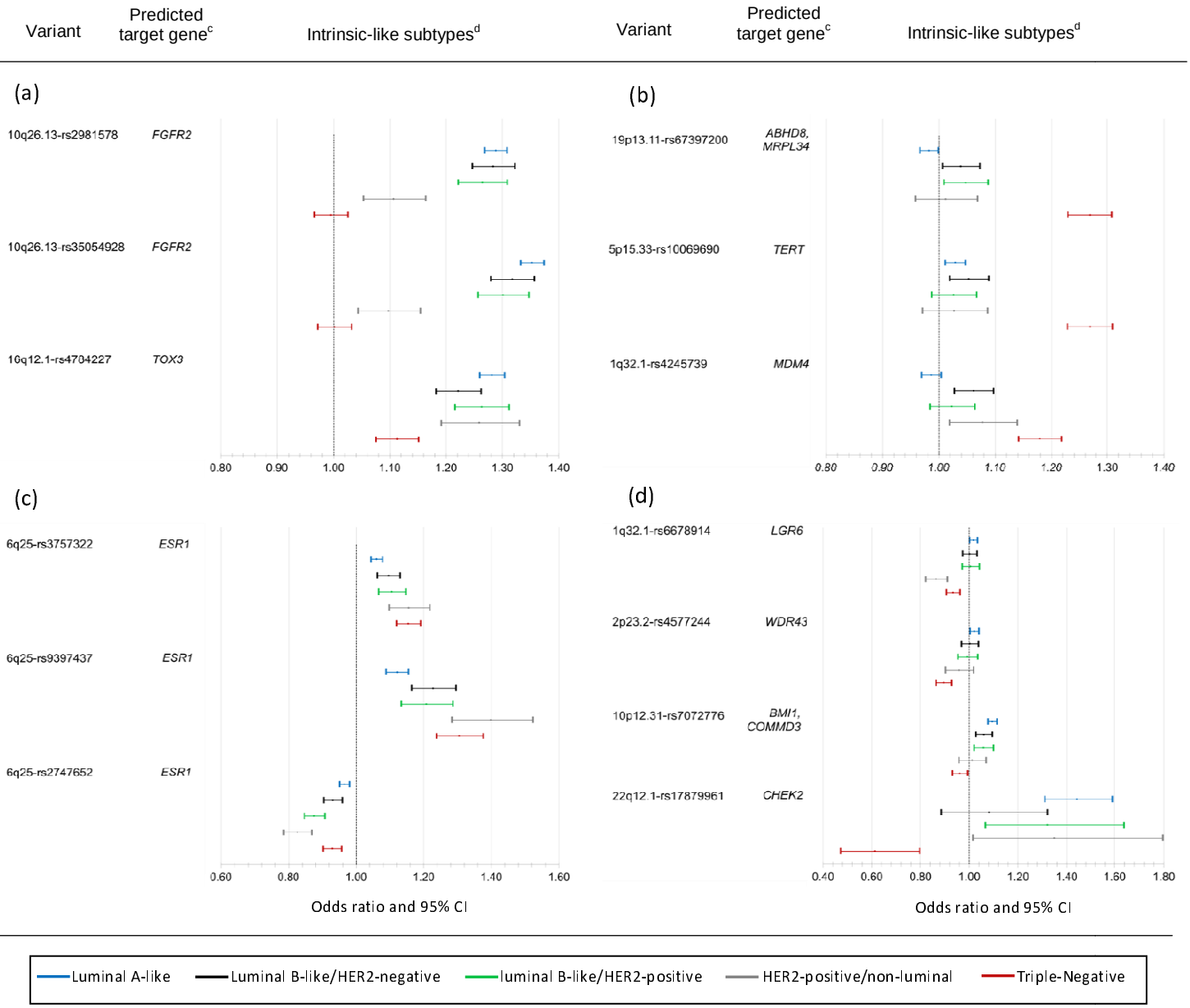


Figure 3. Results from fixed-effects two-stage polytomous models for risk associations[a] with intrinsic-like subtypes (model 3) for variants with evidence of heterogeneity by tumor markers in the two-stage model (model1)[b]; panels show examples of variants (a) most strongly associated with luminal-like subtypes, (b) most strongly associated with TN subtypes, (c) associated with all subtypes with varying strengths of association, and (d) associated with luminal A-like and TN subtypes in different directions. See **Supplementary Figure 3** for more details.



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

[b] Model 1, mixed-effects two-stage polytomous model testing for global heterogeneity according to estrogen receptor (ER), progesterone receptor (PR), huma epidermal growth factor receptor 2 (HER2) and grade

[c] Predicted target genes as reported in Fachal L, et al. Nature genetics 2020; 52 (1), 56-73

[d] Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); Luminal B-like/HER2-negative (ER+ and/or PR+, HER2-, grade 3); luminal B-like/HER2-positive (ER+ and/or PR+, HER2+); HER2-positive/non-luminal (ER- and PR-, HER2+), and triple-negative (ER-, PR-, HER2-)

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] among cases of invasive 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	27,510 (54%)
Luminal B-like/HER2-negative	6,804 (13%)
Luminal B-like/HER2-positive	6,511 (13%)
HER2-positive/non-luminal	2,797 (6%)
Triple-negative	7,178 (14%)
Unknown	55,771

a Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); Luminal B-like/HER2-negative (ER+ and/or PR+, HER2-, grade 3); Luminal B-like/HER2-positive (ER+ and/or PR+, HER2+); HER2-positive/non-luminal (ER- and PR-, HER2+), and triple-negative (ER-, PR-, HER2-)