**Common breast cancer risk loci predispose to distinct tumor subtypes**

Thomas U. Ahearn1\*, Haoyu Zhang2,1\*, Kyriaki Michailidou3,4, Roger L. Milne5,6,7, Manjeet K. Bolla3, Joe Dennis3, Michael Lush3, Qin Wang3, Irene L. Andrulis8,9, Hoda Anton-Culver10, Kristan J. Aronson11, Paul L. Auer12,13, Javier Benitez14,15, Stig E. Bojesen16,17,18, Hiltrud Brauch19,20,21, Hermann Brenner22,23,21, Barbara Burwinkel24,25, Federico Canzian26, Jose E. Castelao27, Jenny Chang-Claude28,29, Georgia Chenevix-Trench30, Christine L. Clarke31, Angela Cox32, Kamila Czene33, Mary B. Daly34, Peter Devilee35,36, Thilo Dörk37, Alison M. Dunning38, Miriam Dwek39, Diana M. Eccles40, D.Gareth Evans41,42, Peter A. Fasching43,44, Lin Fritschi45, Manuela Gago-Dominguez46,47, Susan M. Gapstur48, José A. García-Sáenz49, Mia M. Gaudet48, Graham G. Giles50,51,52, Mark S. Goldberg53,54, David E. Goldgar55, Anna González-Neira15, Pascal Guénel56, Christopher A. Haiman57, Per Hall33,58, Ute Hamann59, Antoinette Hollestelle60, Maartje J. Hooning60, John L. Hopper51, Milena Jakimovska61, Anna Jakubowska62,63, Esther M. John64, Rudolf Kaaks28, Cari M. Kitahara65, Stella Koutros1, Vessela N. Kristensen66,67, Diether Lambrechts68,69, Annika Lindblom70,71, Arto Mannermaa72,73,74, Sara Margolin58,75, Alfons Meindl76, Usha Menon77, Heli Nevanlinna78, Kenneth Offit79,80, Håkan Olsson81, Julian Peto82, Dijana Plaseska-Karanfilska61, Ross Prentice12, Paolo Radice83, Gad Rennert84, Atocha Romero85, Emmanouil Saloustros86, Dale P. Sandler87, Elinor J. Sawyer88, Rita K. Schmutzler89,90, Xiao-Ou Shu91, Melissa C. Southey92,93, John J. Spinelli94,95, Jennifer Stone96,51, Anthony J. Swerdlow97,98, Rulla M. Tamimi99,100,101, William J. Tapper102, Jack A. Taylor87,103, MaryBeth Terry104, Melissa A. Troester105, Celine M. Vachon106, Clarice R. Weinberg107, Robert Winqvist108,109, Alicja Wolk110,111, Wei Zheng91, Fergus J. Couch112, Jacques Simard113, Peter Kraft101,100, Douglas F. Easton38,3, Paul D.P. Pharoah38,3, Marjanka K. Schmidt114,115, Montserrat García-Closas1,116\*\*, Nilanjan Chatterjee2,117\*\*, on behalf of the Breast Cancer Association Consortium

\*Contributed equally

\*\* Contributed equally

**Word count abstract**: 250/250

**Word count main text**: 3,000/3,000

**Conflicts of interest:** None to report

**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](mailto:montserrat.garcia-closas@nih.gov)

1Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, USA, 2Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA, 3Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, 4Department of Electron Microscopy/Molecular Pathology and The Cyprus School of Molecular Medicine, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus, 5Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Victoria, Australia, 6Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia, 7Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia, 8Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Canada, 9Department of Molecular Genetics, University of Toronto, Toronto, Canada, 10Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, USA, 11Department of Public Health Sciences, and Cancer Research Institute, Queen’s University, Kingston, Canada, 12Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, USA, 13Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, USA, 14Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain, 15Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, 16Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark, 17Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark, 18Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, 19Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany, 20University of Tübingen, Tübingen, Germany, 21German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany, 22Division of Clinical Epidemiology and Aging Research, C070, German Cancer Research Center (DKFZ), Heidelberg, Germany, 23Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany, 24Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, Germany, 25Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg, University of Heidelberg, Heidelberg, Germany, 26Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany, 27Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain, 28Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, 29Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany, 30Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia, 31Westmead Institute for Medical Research, University of Sydney, Sydney, Australia, 32Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK, 33Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, 34Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, USA, 35Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands, 36Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands, 37Gynaecology Research Unit, Hannover Medical School, Hannover, Germany, 38Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK, 39Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK, 40Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK, 41Division 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, 42Manchester 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, 43David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, USA, 44Department of Gynecology and Obstetrics, Comprehensive Cancer Center ER-EMN, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany, 45School of Public Health, Curtin University, Perth, Australia, 46Genomic 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, 47Moores Cancer Center, University of California San Diego, La Jolla, USA, 48Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, USA, 49Medical Oncology Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain, 50Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Australia, 51Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia, 52Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia, 53Department of Medicine, McGill University, Montréal, Canada, 54Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, Canada, 55Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, USA, 56Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France, 57Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA, 58Department of Oncology, Södersjukhuset, Stockholm, Sweden, 59Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany, 60Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands, 61Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia, 62Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland, 63Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland, 64Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, USA, 65Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, USA, 66Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway, 67Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, 68VIB Center for Cancer Biology, VIB, Leuven, Belgium, 69Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium, 70Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden, 71Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden, 72Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland, 73Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland, 74Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland, 75Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden, 76Department of Gynecology and Obstetrics, Ludwig Maximilian University of Munich, Munich, Germany, 77MRC Clinical Trials Unit at UCL, Institute of Clinical Trials & Methodology, University College London, London, UK, 78Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, 79Clinical Genetics Research Lab, Department of Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, USA, 80Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, USA, 81Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden, 82Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK, 83Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy, 84Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel, 85Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain, 86Department of Oncology, University Hospital of Larissa, Larissa, Greece, 87Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, USA, 88Research Oncology, Guy’s Hospital, King&apos;s College London, London, UK, 89Center for Hereditary Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, 90Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, 91Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, USA, 92Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Australia, 93Department of Clinical Pathology, The University of Melbourne, Melbourne, Australia, 94Population Oncology, BC Cancer, Vancouver, Canada, 95School of Population and Public Health, University of British Columbia, Vancouver, Canada, 96The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Australia, 97Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK, 98Division of Breast Cancer Research, The Institute of Cancer Research, London, UK, 99Channing Division of Network Medicine, Department of Medicine, Brigham and Women&apos;s Hospital and Harvard Medical School, Boston, USA, 100Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA, 101Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, USA, 102Faculty of Medicine, University of Southampton, Southampton, UK, 103Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, USA, 104Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA, 105Department of Epidemiology, Gilliungs School of Global Public Health and UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, USA, 106Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, USA, 107Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, USA, 108Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland, 109Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland, 110Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, 111Department of Surgical Sciences, Uppsala University, Uppsala, Sweden, 112Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, USA, 113Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval, Research Center, Québec City, Canada, 114Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, 115Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands, 116Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK, 117Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, USA

# **Abstract**

**Background**: Genome-wide association studies have identified over 170 common breast cancer susceptibility loci, many of them with differential associations by estrogen receptor (ER). How these variants are related to other tumor features is unclear.

**Methods**: Analyses included 106,571 invasive breast cancer cases and 95,762 controls of European ancestry with data on 178 genotyped or imputed single nucleotide polymorphisms (SNPs). We used two-stage polytomous logistic regression models to evaluate SNPs in relation to multiple tumor features (ER, progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and grade) adjusting for each other, and in relation to molecular subtypes.

**Results**: Nearly half of the SNPs (85 out of 178) were associated with at least one tumor feature (false discovery rate <5%). Case-case comparisons identified ER and grade as the most common heterogeneity sources, followed by PR and HER2. Case-control comparisons among these 85 SNPs with molecular subtypes identified four main clusters of SNPs: two clusters most strongly or exclusively (P<0.05) associated with luminal-like subtypes (65 SNPs), one cluster associated with all subtypes at various strengths (5 SNPs), and one cluster primarily associated with non-luminal tumors, especially triple-negative (TN) disease (15 SNPs). A *CHEK2* variant (rs17879961) showed significant risk associations with luminal A-like (P=9.26x10-14) and TN (P=2.55x10-4) subtypes in opposite directions.

**Conclusion**: Breast cancer susceptibility loci have complex associations with multiple tumor features. ER and tumor grade are the most common sources of heterogeneity. These findings provide insights into the genetic predisposition of breast cancer subtypes and can inform subtype-specific risk predictions.

**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 molecular subtypes (i.e. luminal A-like, luminal B, HER2-negative like, luminal B-like, HER2-enriched like, and basal-like/triple negative) that are correlated with intrinsic subtypes defined by gene expression panels [2]. 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 breast cancer subtype [3, 4].

Breast cancer genome-wide association studies (GWAS) have identified over 170 susceptibility single nucleotide polymorphisms (SNPs), many of which are differentially associated with ER-positive than ER-negative disease [5]. These include 20 SNPs that primarily predispose to ER-negative or TN disease [6, 7]. However, few studies have evaluated associations of SNPs with other tumor features, or simultaneously studied multiple, correlated tumor markers to identify source(s) of etiologic heterogeneity [6, 8-12]. We recently developed a two-stage polychotomous logistic regression that efficiently characterizes etiologic heterogeneity while accounting for tumor marker correlations and missing tumor data [13]. This method can help characterize complex relationships between susceptibility variants and multiple tumor features, clarifying breast cancer subtype etiologies and increasing the power to generate more accurate risk estimates between susceptibility variants and uncommon subtypes.

In this report, we applied this novel methodology to a large study population from the Breast Cancer Association Consortium (BCAC) to characterize risk associations of 178 known breast cancer susceptibility SNPs with 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 [5, 6] and in the **Supplemental 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); **Supplemental Table 1**). We evaluated 178 susceptibility variants that were identified or replicated in prior BCAC analyses [5, 6]. Genotyped or imputed SNPs marking the 178 susceptibility loci were determined using the iCOGS and the OncoArray genotyping arrays and imputed to the 1000 Genomes Project (Phase 3) reference panel.

**Statistical Analysis**

The statistical methods, including a detailed discussion of the two-stage polytomous logistic regression, is provided in the **Supplemental Methods** and elsewhere [13].Briefly, we identified SNPs showing heterogeneity evidence by using a mixed-effects two-stage polytomous model to evaluate a global heterogeneity test that evaluates if a SNP’s risk estimates vary by at least one of the underlying tumor characteristics. We accounted for multiple testing of the global heterogeneity test using a false discovery rate (FDR) < 0.05 under the Benjamini-Hochberg procedure [14]. Among SNPs with heterogeneity evidence, we used the fixed-effects two-stage model to evaluate the specific tumor marker heterogeneity test, which identifies the specific tumor marker(s) contributing to observed heterogeneity, adjusting for the other tumor markers in the model. Marker-specific P<0.05 was considered statistically significant. Our primary analysis evaluated for heterogeneity by ER, PR, HER2 and grade. As a secondary analysis we fit an extended model with an additional term for TN status to test for differences between TN vs non-TN subtypes. The two-stage model implements an efficient Expectation-Maximization algorithm [15] to essentially perform iterative “imputation” of missing tumor characteristics. We fit an additional two-stage model to estimate case-control ORs and 95% confidence intervals (CI) between the SNPs and five intrinsic-like subtypes defined by combinations of ER, PR, HER2 and grade (**see Supplemental Methods**): (1) luminal A-like, (2) luminal B, HER2-negative-like, (3) luminal B-like, (4) HER2-enriched-like and (5) TN. For all analyses we analyzed OncoArray and iCOGS array data separately and then meta-analyzed the results. Models were adjusted for the first 10 principal components for ancestry informative SNPs. We used Euclidean distance in cluster analyses to visualize the results to identify common heterogeneity sources.

# **Results**

The mean (SD) age at diagnosis (cases) and enrollment (controls) was 56.6 (12.2) and 56.4 (12.2) years, respectively. Eight-one percent of tumors were ER-positive, 68% PR-positive, 83% HER2-negative and 69% grade 1 or 2 (**Table 1;** **Supplemental Table 1**). The most common molecular 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**).

The two-stage models including terms for ER, PR, HER2 and grade, simultaneously adjusting for each other, identified 85 of 178 SNPs (47.7%) with heterogeneity evidence by at least one tumor feature (FDR<5%). ER and grade most often contributed to observed heterogeneity (45 and 34 SNPs, respectively, had marker-specific P<0.05), and 30 SNPs were significantly associated with more than one tumor characteristic (**Figure 1; Supplementary Figure 1)**. Seventeen of these 85 SNPs showed no associations with any individual tumor marker at P<0.05 in their corresponding fixed-effect two-stage models (**Figure 1** and **Supplemental Figure 1**). Twelve SNPs (**Figure 1; Supplemental Figure** 1) were significantly associated exclusively with grade (1p22.3-rs17426269 (Pgrade=1.52x10-02), 1q21.2-rs12048493 (*OTUD7B*; Pgrade=3.09x10-03), 1q22-rs4971059 (*TRIM46*; Pgrade=1.84x10-02), 3p.24.1-rs12493607 (*TGFBR2*; Pgrade=7.78x10-10), 3q26.31-rs58058861 (Pgrade=1.54x10-03), 5p13.3-rs2012709 (Pgrade=6.25x10-03), 10q22.3-rs704010 (*ZMZ1*; Pgrade=2.87x10-04), 11q24.3-rs11820646 (Pgrade=3.18x10-02), 13q13.1-rs11571833 (*BRCA2*; Pgrade=1.78x10-03), 17q22-rs2787486 (Pgrade=1.70x10-04), 19p13.11-rs4808801 (*ELL*; Pgrade=5.10x10-04), 22q13.1-rs738321 (*PLA2G6*; Pgrade=2.80x10-02)), four SNPs were associated exclusively with PR (5q11.1-rs72749841 (PPR=5.46x10-03), 9q31.2-rs10816625 (PPR=3.70x10-02), 9q31.2-rs10759243 (PPR=2.18x10-05), 10q26.12-rs11199914 (PPR=4.39x10-02)) and one SNP was associated exclusively with HER2 (10q21.2-rs10995201 (ZNF365; PHER2= 2.60x10-02)).

In case-control comparisons for the 85 SNPs with evidence for global heterogeneity, we identified four main clusters of SNPs according to the p-values for risk associations with each subtype (**Figure 2 and Supplemental Figure 2**). Sixty-five SNPs in clusters 1 (n=3 SNPs) and 4 (n=62 SNPs) showed the strongest evidence for associations with risk for luminal-like subtypes, cluster 2 (n=5 SNPs) was associated with risk for all subtypes at varying strengths and cluster 3 (n=15 SNPs) with stronger evidence for TN or non-luminal subtype associations. **Supplemental Table 2** shows the associations between all 178 SNPs and the intrinsic-like subtypes.

Cluster 1 included two correlated (r2=0.73) SNPs at 10q26.13-rs2981578 and 10q26.13-rs35054928 (*FGFR2*) that were strongly associated with risk of all the luminal-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 1.37), P=5.00x10-300 for luminal A-like, respectively), weakly associated with risk of HER2-enriched-like subtype (OR (95%CI)=1.11 (1.05 to1.16), P=6.32x10-05 and OR (95% CI)=1.10 (1.04 to 1.15), P=3.07x10-4, respectively), but not significantly associated (P>0.05) with TN tumors (**Figures 2-3 and Supplemental Figure 2**). Case-case comparisons showed the strongest evidence for associations with ER (PER=1.27x10-30 for rs2981578 and PER=9.98x10-38 for rs35054928 (**Figure 1; Supplemental Figure** 1). In the extended two-stage model that additionally included a term for TN status (**Supplemental Figure 3**), both SNPs were associated with TN status (PTN=3.79x10-06 for rs2981578 and PTN=1.27x10-06 for rs35054928). A third SNP in *FGFR2* (10q26.13-rs45631563) showed similar patterns of associations, but fell into cluster 4 since, unlike rs2981578 and rs2981578, it was not significantly associated with risk of the HER2-enriched-like subtype (**Figures 2-3 and Supplemental Figure 2**). Case-case comparisons for rs45631563 showed associations with ER (PER=9.09x10-7) and grade (Pgrade=2.56x10-3). A third SNP in cluster 1, 16q12.1-rs4784227 (*TOX3*), was most strongly associated with luminal-like subtypes (e.g. OR (95%CI)=1.28 (1.26 to 1.30), P=6.68x10-176 for luminal A-like), and the HER2-enriched-like tumors (OR (95%CI)=1.26 (1.19 to 1.33), P=3.04x10-16), and weaker so with TN tumors (OR (95%CI)=1.11 (1.08 to 1.15), P=6.05x10-10; **Figures 2-3 and, Supplemental Figure 2**). In case-case analyses, rs4784227 was significantly associated with all tumor markers, particularly PR (PPR=2.16x10-4; **Figure 1, Supplemental Figure 1**).

Five SNPs in cluster 2 were associated, to different extents, with all five intrinsic-like subtypes. 6q25-rs2747652 (*ESR1*) and 1p36.22-rs616488 (*PEX14*) were associated particularly with risk of HER2-positive subtypes (**Figures 2-3**; **Supplemental Figures 3**). In case-case comparisons these SNPs were associated with HER2 status (PHER2=1.84x10-7 for rs2747652 and PHER2=2.51x10-6 for rs616488), and grade (PGrade=3.82x10-5 for rs2747652 and PGrade=0.02 for rs616488) (**Figure 1 and Supplemental Figure 1**). Two additional SNPs in *ESR1* (6q25-rs9397437 and 6q25-rs3757322) were most strongly associated with risk of TN disease (**Figures 2-3**; **Supplemental Figure 2**), and in case-case comparisons were associated with ER (PER=4.72x10-3 for rs9397437 and PER=3.64x10-2 for rs3757322) and grade (Pgrade=2.87x10-5 for rs9397437 and Pgrade=2.34x10-3 for rs3757322; **Figure 1 and Supplemental Figure 1**). 13q13.1-rs11571833 (*BRCA2*) was associated with risk of all subtypes, but case-case comparisons showed an association only with grade (PGrade=1.78x10-3). In case-control comparisons the ORs and 95% CIs for rs11571833 with 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 1.35), P=5.0x10-12; and 1.08 (0.97-1.20), P=0.15, respectively (**Supplemental Figure 5**).

Cluster 3 included 15 SNPs most strongly associated with risk of HR-negative subtypes, including three SNPs with the strongest evidence for associations with TN disease: 19p13.11-rs67397200 (OR (95% CI)=1.27 (1.23 to 1.31), P=1.07x10-50),5p15.33-rs10069690 (*TERT*, OR (95% CI)=1.27 (1.23 to 1.31), P=3.79x10-48) and 1q32.11­-rs4245739 (*MDM4*, OR (95% CI)=1.18 (1.14to 1.22), P=2.72x10-23). Two SNPs at 11q22.3 (*KDELC2*), rs11374964 (OR (95% CI)=0.90 (0.88-0.93), P=2.71x10-11) and rs74911261 (OR (95% CI)=0.93 (0.90-0.96), P=2.71x10-11), were exclusively associated with TN disease (**Figures 2-3; Supplemental Figure 2**). In the extended model these five SNPs were associated with TN status (PTN<0.05; Supplemental Figure 3). The remaining 10 SNPs were all associated with TN disease (P­TN<0.05), and in case-case comparisons seven of these 10 SNPs were exclusively associated with ER (PER<0.05): 1q32.1-rs6678914 (*LGR6*), 2p23.2-rs4577244 (*WDR43*), 8p23.3-rs66823261 (*RPL23AP53*), 13q22.1-rs6562760, 16q12.2-rs11075995 (*FTO*), 16p13.3-rs11076805 (*ADCY9*) and 18q12.1-rs36194942 (*CDH2*)). 5p15.33-rs3215401 (*TERT*) was associated with both ER (PER=2.22x10-03) and PR (PPR=4.65x10-02). 2p24.1-rs12710696 and 19q12-rs113701136 (*CCNE1*) showed no associations with any of the individual tumor markers (**Figures 2-3**; **Supplemental Figure 2**). Besides rs11374964 and rs74911261, SNPs in this cluster were not HR-negative or TN-specific SNPs as they were also associated with luminal-like subtypes. Three SNPs showed weak associations with luminal A-like disease (OR (95% 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 rs6678914; and OR (95% CI)=1.02 (1.00 to 1.04); P=0.02 for rs4577244) in an opposite direction of their associations with TN disease (OR (95% CI)=0.93 (0.91 to 0.96); P=1.07x10-4 for rs6678914 and OR (95% CI)=0.90 (0.86 to 0.93); P=2.99x10-9 for rs4577244; **Figures 2-3**; **Supplemental Figure 2**).

Cluster 4 (n=62 SNPs) showed evidence for being associated with risk of luminal-subtypes, especially luminal A-like disease. The five SNPs in this cluster showing the strongest evidence of being associated with risk of the luminal A-like subtype were: 5q11.2-rs62355902 (*MAP3K1,* OR (95% CI)=1.22 (1.20 to 1.25), P=3.94X10-85), 11q13.3-rs75915166 (*CCND1*, OR (95% CI)=1.44 (1.40 to 1.48), P=2.26X10-131), 11q13.3-rs554219 (*CCND1*, OR (95% CI)=1.33 (1.30 to 1.37), P=1.31X10-98), 1p11.2-rs11249433 (*EMBP1*, OR (95% CI)=1.17 (1.15 to 1.19), P=5.25X10-90) and 5p12-rs10941679 (OR (95% CI)=1.20 (1.18 to 1.22), P=1.39X10-93; **Figure 2 and Supplemental Figures 2,4**). In case-case comparisons these five SNPs were associated with grade (Pgrade=2.32x10-02 for rs62355902, Pgrade=1.74x10-13 for rs75915166, Pgrade=2.14x10-12 for rs554219, Pgrade=2.20x10-12 for rs11249433 and Pgrade=2.47x10-06 for rs10941679) and with at least one other tumor marker. Eighteen of the 62 SNPs also showed weaker evidence for being associated with non-luminal subtypes, 11 of which were associated with risk of TN disease (PTN<0.05): 5q11.2-rs62355902, 5p12-rs10941679, 12q22-rs17356907 (*NTN4*), 10q21.2-rs10995201 (*ZNF365*), 10p12.31-rs7072776 (*DNAJC1*), 19p13.11-rs4808801 (*ELL*), 10p12.31-rs11814448 (*DNAJC1*), 8q21.11-rs6472903, 11q24.3-rs11820646, 19p13.13-rs78269692 (*NFIX1*) and 22q12.1-rs17879961 (*CHEK2*; **Figure 2 and Supplemental Figures 2,4**). Notably, rs11814448 and rs17879961 were associated with luminal A-like disease (OR (95% CI)=1.10 (1.08 to 1.11); P=4.96x10-27 for rs11814448; and OR (95% CI)=1.44 (1.31 to 1.59); P=9.26x10-14 for rs17879961) and TN disease (OR (95% CI)=0.61 (0.47 to 0.80); P=2.55x10-4 for rs11814448 and OR (95% CI)=0.61 (0.47 to 0.80); P=2.55x10-4 for rs17879961) in opposite directions. In case-case comparisons rs11814448 showed no associations (P>0.05) with any individual tumor marker, and rs17879961 was associated with the TN phenotype in the extended model (PTN=1.45x10-5; **Supplemental Figure 3**).

**Supplemental Figure 5** shows case-control associations by tumor grade for the 12 SNPs significantly associated exclusively with grade in case-case comparisons. rs11571833, rs17426269 and rs11820646 showed stronger evidence for predisposing to risk of high-grade subtypes, and the remaining SNPs showed stronger evidence for predisposing to risk of low-grade subtypes.

# **Discussion**

We found compelling evidence that about half of the investigated breast cancer susceptibility loci (85 of 178 SNPs) predispose to tumors of 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 molecular subtypes defined by their combinations.

Many of the heterogenous SNPs predisposed to risk of all subtypes, but with different magnitudes. For example, 21 of 65 SNPs found predominately associated with risk of luminal-like subtypes were also associated with risk of at least one of the non-luminal subtypes. These include SNPs identified in early GWAS for overall breast cancer [16], such as *FGFR2* loci (rs35054928 and rs2981578)[17, 18], 8q24.21 (rs13281615) [17] and *IGFBP5* (rs4442975) [17] that were associated with luminal-like and HER2-enriched-like subtypes. rs4784227 located near *TOX3* [17, 19] and rs62355902 located in a *MAP3K1* [17] regulatory element, were associated with risk of all five subtypes. Of the five SNPs found significantly 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 [6]. rs17879961, a likely causal [20] mis-sense variant located in a *CHEK2* functional domain that reduces or abolishes substrate binding [21], 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 [22, 23]. 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 SNPs that predispose to ER-negative disease, of which five SNPs were only or most strongly associated with risk of TN disease (rs4245739, rs10069690, rs74911261, rs11374964, and rs67397200) [6, 7]. We further confirmed these five SNPs to be most strongly associated with TN disease and found these SNPs associated with TN status in the extended model. The remaining previously identified 15 SNPs all showed associations with risk of HR-negative disease, and all but four SNPs (rs17350191, rs200648189, rs6569648, and rs322144) were found with evidence of global heterogeneity. Among the SNPs in cluster 3, rs3215401 was the only SNP that was not identified in a prior ER-negative GWAS [6, 7]. rs3215401 was identified in a fine-mapping analysis of *TERT* and, consistent with our findings, reported to be most strongly associated with ER-negative disease but was also associated with ER-positive disease [24].

Little is known regarding PR and HER2 as sources of etiologic heterogeneity independent of ER or TN status. Of the four SNPs significantly associated only with PR, rs10759243 [5, 25], rs11199914 [5] and rs72749841 [5] were previously found primarily associated with risk of ER-positive disease, and rs10816625 was found associated with risk of ER-positive/PR-positive tumors, but not other ER/PR combinations [11]. rs10995201 was the only variant found to be solely associated with HER2 status, although the evidence was not strong and requires further confirmation. Previously rs10995201 showed no evidence of being associated with ER tumor status [26]. Among all SNPs found associated with PR or HER2, few have been investigated for heterogeneity according to PR or HER2 while adjusting for ER [8-12]. We previously reported rs10941679 to be associated with PR-status, independent of ER, and also with grade [9]. We also found suggestive evidence of PR-specific heterogeneity for 16q12-rs3803662 [12], which is in high LD (r2= 0.77) with rs4784227 (*TOX3*), which was 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-enriched subtype and high grade independent of ER [8]. rs2747652 overlaps an enhancer region and is associated with reduced *ESR1* and *CCDC170* expression [8].

Histologic grade is a composite of multiple tumor characteristics including the mitotic count, nuclear pleomorphism, and the degree of tubule or gland formation [27]. Among the 12 SNPs identified with only evidence of grade heterogeneity, rs17426269, rs11820646, and rs11571833 were found most strongly associated with grade 3 disease. rs11571833 lies in a *BRCA2* coding region and produces a truncated form of the protein [28] 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 [29]. To our knowledge, rs17426269 and rs11820646 have not been investigated in relation to grade heterogeneity. The remaining 9 SNPs were all more strongly associated with grade 1 or grade 2 disease. Five of these SNPs were previously reported to be associated primarily with ER-positive disease [5, 16, 30], highlighting the importance of accounting for multiple tumor characteristics to better illuminate complex heterogeneity sources.

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 novel two-stage polychotomous logistic regression enabled adjusting for multiple, correlated tumor markers and accounted for missing tumor marker data. This is a more powerful and efficient modeling strategy for identifying heterogeneity sources among highly correlated tumor markers, compared to standard polytomous logistic regression. However, we did identify 17 SNPs with evidence of heterogeneity for which we did not identify specific tumor characteristic(s) contributing to observed heterogeneity. This is likely explained by the fact that the fixed-effects models evaluating specific tumor markers as heterogeneity sources was less statistically powerful compared to the mixed-effects models that evaluated for evidence of global heterogeneity [13]. Our study was limited by investigating only ER, PR, HER2, and grade as heterogeneity sources and future studies with more detailed tumor characterization could reveal additional sources of etiologic heterogeneity.

In summary, our results provide insights into complex etiologic heterogeneity patterns of common breast cancer susceptibility loci. These findings can inform 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 molecular subtypes that could improve the discriminatory accuracy of subtype-specific polygenic risk scores [31].

**Refences**

1. The Cancer Genome Atlas N, Koboldt DC, Fulton RS*, et al.* Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61.

2. Curigliano G, Burstein HJ, E PW*, 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(8):1700-1712.

3. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. Biochim Biophys Acta 2015;1856(1):73-85.

4. Yang XR, Chang-Claude J, Goode EL*, et al.* Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst 2011;103(3):250-63.

5. Michailidou K, Lindstrom S, Dennis J*, et al.* Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551(7678):92-94.

6. Milne RL, Kuchenbaecker KB, Michailidou K*, et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 2017;49(12):1767-1778.

7. Garcia-Closas M, Couch FJ, Lindstrom S*, et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet 2013;45(4):392-8, 398e1-2.

8. Dunning AM, Michailidou K, Kuchenbaecker KB*, et al.* Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet 2016;48(4):374-86.

9. Milne RL, Goode EL, Garcia-Closas M*, et al.* Confirmation of 5p12 as a susceptibility locus for progesterone-receptor-positive, lower grade breast cancer. Cancer Epidemiol Biomarkers Prev 2011;20(10):2222-31.

10. Figueroa JD, Garcia-Closas M, Humphreys M*, 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. Hum Mol Genet 2011;20(23):4693-706.

11. Orr N, Dudbridge F, Dryden N*, et al.* Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Hum Mol Genet 2015;24(10):2966-84.

12. Broeks A, Schmidt MK, Sherman ME*, et al.* Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet 2011;20(16):3289-303.

13. Zhang H, Zhao N, Ahearn TU*, et al.* A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. bioRxiv 2018; 10.1101/446039:446039.

14. 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(1):289-300.

15. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via Em Algorithm. Journal of the Royal Statistical Society Series B-Methodological 1977;39(1):1-38.

16. Michailidou K, Hall P, Gonzalez-Neira A*, et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013;45(4):353-61, 361e1-2.

17. Easton DF, Pooley KA, Dunning AM*, et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447(7148):1087-93.

18. Hunter DJ, Kraft P, Jacobs KB*, et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39(7):870-4.

19. Stacey SN, Manolescu A, Sulem P*, et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007;39(7):865-9.

20. Fachal L, Aschard H, Beesley J*, et al.* Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. bioRxiv 2019; 10.1101/521054:521054.

21. Li J, Williams BL, Haire LF*, et al.* Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. Mol Cell 2002;9(5):1045-54.

22. Wang Y, McKay JD, Rafnar T*, et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 2014;46(7):736-41.

23. McKay JD, Hung RJ, Han Y*, et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 2017;49(7):1126-1132.

24. Bojesen SE, Pooley KA, Johnatty SE*, et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet 2013;45(4):371-84, 384e1-2.

25. Li X, Zou W, Liu M*, et al.* Association of multiple genetic variants with breast cancer susceptibility in the Han Chinese population. Oncotarget 2016;7(51):85483-85491.

26. Darabi H, McCue K, Beesley J*, et al.* Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. Am J Hum Genet 2015;97(1):22-34.

27. 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(5):403-10.

28. Mazoyer S, Dunning AM, Serova O*, et al.* A polymorphic stop codon in BRCA2. Nat Genet 1996;14(3):253-4.

29. Meeks HD, Song H, Michailidou K*, et al.* BRCA2 Polymorphic Stop Codon K3326X and the Risk of Breast, Prostate, and Ovarian Cancers. J Natl Cancer Inst 2016;108(2).

30. Darabi H, Beesley J, Droit A*, et al.* Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Sci Rep 2016;6:32512.

31. Mavaddat N, Michailidou K, Dennis J*, et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019;104(1):21-34.