Open access • Posted Content • DOI:10.1101/778605

Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses — Source link

Haoyu Zhang, Haoyu Zhang, Thomas U. Ahearn, Julie Lecarpentier ...+309 more authors

Institutions: Johns Hopkins University, National Institutes of Health, University of Cambridge, QIMR Berghofer Medical Research Institute ...+120 more institutions

Published on: 24 Sep 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Breast cancer, Estrogen receptor, Progesterone receptor and Genome-wide association study

---

Article:
Zhang, H, Ahearn, TU, Lecarpentier, J et al. (269 more authors) (2020) Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nature Genetics, 52 (6). pp. 572-581. ISSN 1061-4036

https://doi.org/10.1038/s41588-020-0609-2

Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses

Haoyu Zhang[1,2*], Thomas U. Ahearn[1*], Julie Lecarpentier[3], Daniel Barnes[3], Jonathan Beesley[4], Guanghao Qi[2], Xia Jiang[5], Tracy O'Mara[4], Ni Zhao[2], Manjeet K. Bolla[6], Alison M. Dunning[3], Joe Dennis[6], Qin Wang[6], Kristiina Aittomäki[7], Irene L. Andrulis[8], Hoda Anton-Culver[9], Volker Arndt[10], Kristan J. Aronson[11], Banu K. Arun[12], Paul L. Auer[13,14], Jacopo Azzollini[15], Daniel Barrowdale[16], Heiko Becher[17], Matthias W. Beckmann[18], Sabine Behrens[19], Javier Benitez[20], Katarzyna Bialkowska[21], Ana Blanco[22,23,24], Carl Blomqvist[25,26], Stig E. Bojesen[27,28,29,30], Bernardo Bonanni[31], Davide Bondavalli[31], Ake Borg[32], Hiltrud Brauch[33,34,35], Hermann Brenner[35,36,37], Ignacio Briceno[38], Annegien Broeks[39], Sara Y. Brucker[40], Thomas Brüning[41], Barbara Burwinkel[42,43], Saundra S. Buys[44], Helen Byers[45], Trinidad Caldés[46], Maria A. Caligo[47], Mariarosaria Calvello[31], Federico Canzian[48], Jose E. Castelao[49], Jenny Chang-Claude[19,50], Stephen J. Chanock[1], Melissa Christiaens[51], Wendy K. Chung[52], Kathleen B.M. Claes[53], Christine L. Clarke[54], Sten Cornelissen[39], Fergus J. Couch[55], Angela Cox[56], Simon S. Cross[57], Kamila Czene[58], Mary B. Daly[59], Peter Devilee[60], Orland Diez[61], Susan M. Domchek[62], Thilo Dörk[63], Miriam Dwek[64], Diana M. Eccles[65], Arif B. Ekici[66], D.Gareth Evans[45,67], Peter A. Fasching[18,68], Jonine Figueroa[69], Lenka Foretova[70], Florentia Fostira[71], Eitan Friedman[72], Debra Frost[16], Manuela Gago-Dominguez[73,74], Susan M. Gapstur[75], Judy Garber[76], José A. García-Sáenz[46], Mia M. Gaudet[75], Simon A. Gayther[77], Graham G. Giles[78,79,80], Andrew K. Godwin[81], Mark S. Goldberg[82,83,84], David E. Goldgar[85], Anna González-Neira[30], Mark H. Greene[86], Jacek Gronwald[21], Pascal Guénel[87], Lothar

Häberle[88], Eric Hahnen[89], Christopher A. Haiman[90], Christopher R. Hake[91], Per Hall[58,92], Ute Hamann[93], Elaine F. Harkness[94,95], Frans B.L. Hogervorst[96], Bernd Holleczek[97], Antoinette Hollestelle[98], Maartje J. Hooning[98], Robert N. Hoover[1], John L. Hopper[79], Anthony Howell[99], Hanna Hübner[18], Peter J. Hulick[100], Evgeny N. Imyanitov[101], kConFab Investigators[102,103], Claudine Isaacs[104], Louise Izatt[105], Milena Jakimovska[106], Anna Jakubowska[21,107], Paul James[108], Ramunas Janavicius[109], Wolfgang Janni[110], Esther M. John[111], Michael E. Jones[112], Audrey Jung[19], Rudolf Kaaks[19], Pooja M Kapoor[19], Beth Y. Karlan[113], Renske Keeman[39], Sofia Khan[114], Cari M. Kitahara[115], Yon-Dschun Ko[116], Irene Konstantopoulou[71], Linetta B. Koppert[117], Stella Koutros[1], Vessela N. Kristensen[118,119], Anne-Vibeke Laenkholm[120], Diether Lambrechts[121,122], Susanna C. Larsson[123,124], Pierre Laurent-Puig[125], Conxi Lazaro[126], Emilija Lazarova[127], Fabienne Lesueur[128], Annika Lindblom[129,130], Jolanta Lissowska[131], Wing-Yee Lo[33,132], Jennifer T. Loud[86], Jan Lubinski[21], Alicja Lukomska[21], Robert J. MacInnis[79,133], Arto Mannermaa[134,135,136], Mehdi Manoochehri[93], Siranoush Manoukian[15], Sara Margolin[92,137], Maria Elena Martinez[74,138], Laura Matricardi[139], Catriona McLean[140], Noura Mebirouk[141], Alfons Meindl[142], Usha Menon[143], Austin Miller[144], Marco Montagna[139], Anna Marie Mulligan[145,146], Claire Mulot[125], Taru A. Muranen[114], Katherine L. Nathanson[62], Susan L. Neuhausen[147], Heli Nevanlinna[114], Patrick Neven[51], William G. Newman[45,67], Finn C. Nielsen[148], Liene Nikitina-Zake[149], Jesse Nodora[150,151], Kenneth Offit[152], Edith Olah[153], Olufunmilayo I. Olopade[154,155], Håkan Olsson[156,157], Nick Orr[158], Laura Papi[159], Bernard Peissel[15], Ana Peixoto[160], Beth Peshkin[161], Paolo Peterlongo[162], Julian Peto[6,163], Kelly-Anne Phillips[79,164,165,166], Marion Piedmonte[144], Dijana Plaseska-Karanfilska[106], Karolina Prajzendanc[21], Ross Prentice[13], Brigitte

Rack[110], Paolo Radice[167], Susan J. Ramus[168,169,170], Johanna Rantala[171], Muhammad U. Rashid[93,172], Gad Rennert[173], Harvey A. Risch[174], Atocha Romero[175,176], Matti A. Rookus[177], Matthias Rübner[88], Thomas Rüdiger[178], Emmanouil Saloustros[179], Sarah Sampson[180], Dale P. Sandler[181], Elinor J. Sawyer[182], Maren T. Scheuner[183], Rita K. Schmutzler[184], Andreas Schneeweiss[43,185], Minouk J. Schoemaker[112], Ben Schöttker[10], Leigha Senter[186], Priyanka Sharma[187], Mark E. Sherman[188], Xiao-Ou Shu[189], Christian F. Singer[190], Snezhana Smichkoska[127], Penny Soucy[191], Melissa C. Southey[80], John J. Spinelli[192,193], Jennifer Stone[79,194], Dominique Stoppa-Lyonnet[195], EMBRACE Study[16], GEMO Study Collaborators[141], Anthony J. Swerdlow[112,196], Csilla I. Szabo[197], Rulla M. Tamimi[5,198,199], William J. Tapper[200], Jack A. Taylor[181,201], Manuel R. Teixeira[160,176], MaryBeth Terry[202], Mads Thomassen[203], Darcy L. Thull[204], Marc Tischkowitz[205], Amanda E. Toland[206], Ian Tomlinson[207,208], Diana Torres[38,93], Melissa A. Troester[209], Thérèse Truong[87], Nadine Tung[210], Michael Untch[211], Celine M. Vachon[212], Ans M.W. van den Ouweland[213], Lizet E. van der Kolk[96], Elke M. van Veen[45,67], Elizabeth J. van Rensburg[214], Ana Vega[22,23,24], Barbara Wappenschmidt[184], Clarice R. Weinberg[215], Jeffrey N. Weitzel[216], Hans Wildiers[51], Robert Winqvist[217,218,219,220], Alicja Wolk[107,123,124], Xiaohong R. Yang[1], Drakoulis Yannoukakos[71], Wei Zheng[189], Kristin K. Zorn[221], Monica Zuradelli[222], Roger L. Milne[79,80,223], Peter Kraft[5,199], Jacques Simard[191], Paul D.P. Pharoah[3,6], Kyriaki Michailidou[6,224,225], Antonis C. Antoniou[6], Marjanka K. Schmidt[39,226], Georgia Chenevix-Trench[4**], Douglas F. Easton[3**], Nilanjan Chatterjee[2,227**],Montserrat García-Closas[1**]

[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 Oncology, University of Cambridge, Cambridge, UK, [4]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia, [5]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [6]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [7]Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, [8]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada, [9]Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA, USA, [10]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany, [11]Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON, Canada, [12]Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA, [13]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [14]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA, [15]Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [16]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK, [17]Institute of Medical Biometry

and Epidemiology, University of Hamburg, Hamburg, Germany, [18]Department of Gynecology and Obstetrics, Comprehensive Cancer Center ER-EMN, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, 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]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland, [22]Molecular Medicine Unit, Fundación Pública Galega de Medicina Xenómica, Santiago de Compostela, Spain, [23]Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain, [24]Centro de Investigación en Red de Enfermedades Raras (CIBERER), Santiago de Compostela, Spain, [25]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, [26]Department of Oncology, Örebro University Hospital, Örebro, Sweden, [27]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark, [28]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark, [29]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [30]Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, [31]Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, Milan, Italy, [32]Department of Oncology, Lund University and Skåne University Hospital, Lund, Sweden, [33]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany, [34]iFIT-Cluster of Excellence, University of Tübingen, Tübingen, Germany, [35]German Cancer Consortium

(DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany, [36]Division of Clinical Epidemiology and Aging Research, C070, German Cancer Research Center (DKFZ), Heidelberg, Germany, [37]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany, [38]Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia, [39]Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, [40]Department of Gynecology and Obstetrics, University of Tübingen, Tübingen, Germany, [41]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany, [42]Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, Germany, [43]Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg, University of Heidelberg, Heidelberg, Germany, [44]Department of Medicine, Huntsman Cancer Institute, Salt Lake City, UT, USA, [45]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, [46]Medical Oncology Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain, [47]Section of Molecular Genetics, Dept. of Laboratory Medicine, University Hospital of Pisa, Pisa, Italy, [48]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany, [49]Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain, [50]Cancer Epidemiology Group,

University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [51]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium, [52]Departments of Pediatrics and Medicine, Columbia University, New York, NY, USA, [53]Centre for Medical Genetics, Ghent University, Gent, Belgium, [54]Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia, [55]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA, [56]Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK, [57]Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK, [58]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [59]Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA, [60]Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands, [61]Oncogenetics Group, Vall dHebron Institute of Oncology (VHIO), Barcelona, Spain, [62]Department of Medicine, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA, [63]Gynaecology Research Unit, Hannover Medical School, Hannover, Germany, [64]Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK, [65]Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK, [66]Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany, [67]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, [68]David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, USA, [69]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, UK, [70]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic, [71]Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research "Demokritos", Athens, Greece, [72]The Susanne Levy Gertner Oncogenetics Unit, Chaim Sheba Medical Center, Ramat Gan, Israel, [73]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, [74]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA, [75]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA, [76]Cancer Risk and Prevention Clinic, Dana-Farber Cancer Institute, Boston, MA, USA, [77]Center for Bioinformatics and Functional Genomics and the Cedars Sinai Genomics Core, Cedars-Sinai Medical Center, Los Angeles, CA, USA, [78]Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Victoria, Australia, [79]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia, [80]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia, [81]Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Kansas City, KS, USA, [82]Department of Medicine, McGill University, Montréal, QC, Canada, [83]Division of Clinical Epidemiology, Royal

Victoria Hospital, McGill University, Montréal, QC, Canada, [84]Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia, [85]Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, USA, [86]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, [87]Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France, [88]Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany, [89]Center for Familial Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [90]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [91]Waukesha Memorial Hospital-Pro Health Care, Waukesha, WI, USA, [92]Department of Oncology, Södersjukhuset, Stockholm, Sweden, [93]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany, [94]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK, [95]Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, UK, [96]Family Cancer Clinic, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands, [97]Saarland Cancer Registry, Saarbrücken, Germany, [98]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The

Netherlands, [99]Division of Cancer Sciences, University of Manchester, Manchester, UK, [100]Center for Medical Genetics, NorthShore University HealthSystem, Evanston, IL, USA, [101]N.N. Petrov Institute of Oncology, St. Petersburg, Russia, [102]Peter MacCallum Cancer Center, Melbourne, Victoria, Australia, [103]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia, [104]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA, [105]Clinical Genetics, Guy's and St. Thomas' NHS Foundation Trust, London, UK, [106]Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia, [107]Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland, [108]Parkville Familial Cancer Centre, Peter MacCallum Cancer Center, Melbourne, Victoria, Australia, [109]Hematology, oncology and transfusion medicine center, Dept. of Molecular and Regenerative Medicine, Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania, [110]Department of Gynaecology and Obstetrics, University Hospital Ulm, Ulm, Germany, [111]Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA, [112]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK, [113]David Geffen School of Medicine, Department of Obstetrics and Gynecology, University of California at Los Angeles, Los Angeles, CA, USA, [114]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, [115]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, [116]Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter

Krankenhaus, Bonn, Germany, [117]Department of Surgical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands, [118]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway, [119]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, [120]Depastment of Surgical Pathology, Zealand University Hospital, Slagelse, Denmark, [121]VIB Center for Cancer Biology, VIB, Leuven, Belgium, [122]Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium, [123]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, [124]Department of Surgical Sciences, Uppsala University, Uppsala, Sweden, [125]Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris, France, [126]Molecular Diagnostic Unit, Hereditary Cancer Program, ICO-IDIBELL (Bellvitge Biomedical Research Institute, Catalan Institute of Oncology), CIBERONC, Barcelona, Spain, [127]Ss. Cyril and Methodius University in Skopje, Medical Faculty, University Clinic of Radiotherapy and Oncology, Skopje, Republic of North Macedonia, [128]Genetic Epidemiology of Cancer team, Inserm U900, Paris, France, [129]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden, [130]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden, [131]Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Cancer Center, Oncology Institute, Warsaw, Poland, [132]University of Tübingen, Tübingen, Germany, [133]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia, [134]Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland, [135]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland, [136]Imaging Center,

Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland, [137]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden, [138]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA, [139]Immunology and Molecular Oncology Unit, Veneto Institute of Oncology IOV - IRCCS, Padua, Italy, [140]Department of Anatomical Pathology, The Alfred Hospital, Prahran, Victoria, Australia, [141]Genetic Epidemiology of Cancer team, Inserm U900, Institut Curie, PSL University, Mines ParisTech, Paris, France, [142]Department of Gynecology and Obstetrics, Ludwig Maximilian University of Munich, Munich, Germany, [143]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials & Methodology, University College London, London, UK, [144]NRG Oncology, Statistics and Data Management Center, Roswell Park Cancer Institute, Buffalo, NY, USA, [145]Laboratory Medicine Program, University Health Network, Toronto, ON, Canada, [146]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada, [147]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA, [148]Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark, [149]Latvian Biomedical Research and Study Centre, Riga, Latvia, [150]Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA, [151]Department of Family Medicine and Public Health, School of Medicine, University of California, San Diego, La Jolla, CA, USA, [152]Clinical Genetics Research Lab, Department of Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA, [153]Department of Molecular Genetics, National Institute of Oncology, Budapest, Hungary, [154]Center for Clinical Cancer Genetics, The University of Chicago, Chicago, IL, USA, [155]Department

of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia, [156]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden, [157]Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA, [158]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, Ireland, UK, [159]Unit of Medical Genetics, Department of Biomedical, Experimental and Clinical Sciences,, University of Florence, Florence, Italy, [160]Department of Genetics, Portuguese Oncology Institute, Porto, Portugal, [161]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA, [162]Genome Diagnostics Program, IFOM, The FIRC Institute of Molecular Oncology, Milan, Italy, [163]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK, [164]Department of Medicine, St Vincent's Hospital, The University of Melbourne, Fitzroy, Victoria, Australia, [165]Peter MacCallum Cancer Center, Melbourne, Victoria, Australia, [166]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia, [167]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy, [168]Adult Cancer Program, Lowy Cancer Research Centre, University of NSW Sydney, Sydney, New South Wales, Australia, [169]School of Women's and Children's Health, Faculty of Medicine, University of NSW Sydney, Sydney, New South Wales, Australia, [170]The Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, Australia, [171]Clinical Genetics, Karolinska Institutet, Stockholm, Sweden, [172]Department of Basic Sciences, Shaukat Khanum Memorial Cancer Hospital and Research Centre (SKMCH & RC),

Lahore, Pakistan, [173]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel, [174]Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA, [175]Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain, [176]Biomedical Sciences Institute (ICBAS), University of Porto, Porto, Portugal, [177]Department of Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands, [178]Institute of Pathology, Staedtisches Klinikum Karlsruhe, Karlsruhe, Germany, [179]Department of Oncology, University Hospital of Larissa, Larissa, Greece, [180]Prevent Breast Cancer Centre and Nightingale Breast Screening Centre, Manchester University NHS Foundation Trust, Manchester, UK, [181]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA, [182]Research Oncology, Guy's Hospital, King's College London, London, UK, [183]Cancer Genetics and Prevention Program, University of California San Francisco, San Francisco, CA, USA, [184]Center for Hereditary Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [185]National Center for Tumor Diseases, University Hospital and German Cancer Research Center, Heidelberg, Germany, [186]Clinical Cancer Genetics Program, Division of Human Genetics, Department of Internal Medicine, The Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA, [187]Department of Internal Medicine, Division of Oncology, University of Kansas Medical Center, Westwood, KS, USA, [188]Department of Health Sciences Research, Mayo Clinic College of Medicine, Jacksonville, FL, USA, [189]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville,

TN, USA, [190]Dept of OB/GYN and Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria, [191]Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval, Research Center, Québec City, QC, Canada, [192]Population Oncology, BC Cancer, Vancouver, BC, Canada, [193]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada, [194]The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Western Australia, Australia, [195]Department of Genetics, Inserm U830, Institut Curie, Paris Descartes Sorbonne-Paris-Cité University, Paris, France, [196]Division of Breast Cancer Research, The Institute of Cancer Research, London, UK, [197]National Human Genome Research Institute, National Cancer Institute, Bethesda, MD, USA, [198]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, [199]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [200]Faculty of Medicine, University of Southampton, Southampton, UK, [201]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA, [202]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA, [203]Department of Clinical Genetics, Odense University Hospital, Odence C, Denmark, [204]Department of Medicine, Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA, [205]Program in Cancer Genetics, Departments of Human Genetics and Oncology, McGill University, Montréal, QC, Canada, [206]Department of Cancer Biology and Genetics, The Ohio State University, Columbus, OH, USA, [207]Institute of Cancer and Genomic Sciences, University of Birmingham,

Birmingham, UK, [208]Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK, [209]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, [210]Department of Medical Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA, [211]Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany, [212]Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA, [213]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands, [214]Department of Genetics, University of Pretoria, Arcadia, South Africa, [215]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA, [216]Clinical Cancer Genomics, City of Hope, Duarte, CA, USA, [217]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland, [218]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland, [219]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, [220]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands, [221]Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA, [222]Unità Operativa di Oncologia Medica ed Ematologia, Humanitas Cancer Center Istituto Clinico Humanitas- IRCCS, Milan, Italy, [223]Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne,, Victoria, Australia, [224]Biostatistics Unit, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus, [225]Cyprus School of Molecular Medicine, Nicosia, Cyprus,

[226]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands,

[227]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

*Contributed equally

**Contributed equally

Target journal: Nature Genetics

Word Count for letter:

Introductory paragraph: 183 / the journal suggest having approximately 150

Main text: 1496/1500

Conflicts of interest: None to report

Corresponding Author

Nilanjan Chatterjee

615 N. Wolfe Street

Room E3612

Baltimore, Maryland 21205

nchatte2@jhu.edu

Breast cancer susceptibility variants frequently show heterogeneity in associations by tumor subtype. To identify novel loci, we performed a genome-wide association study (GWAS) including 133,384 breast cancer cases and 113,789 controls, plus 18,908 *BRCA1* mutation carriers (9,414 with breast cancer) of European ancestry, using both standard and novel methodologies that account for underlying tumor heterogeneity by estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status and tumor grade. We identified 32 novel susceptibility loci ($P<5.0x10^{-8}$), 15 of which showed evidence for associations with at least one tumor feature (false discovery rate <0.05). Five loci showed associations ($P<0.05$) in opposite directions between luminal- and non-luminal subtypes. *In-silico* analyses showed these five loci contained cell-specific enhancers that differed between normal luminal and basal mammary cells. The genetic correlations between five intrinsic-like subtypes ranged from 0.49 to 0.87. The proportion of heritability explained by all known susceptibility loci was 31.9% for triple-negative and 45.2% for luminal A-like disease. These findings provide improved understanding of genetic predisposition to breast cancer subtypes and will inform the development of subtype-specific polygenic risk scores.

GWAS have identified over 170 independent breast cancer susceptibility loci, many of which show differential associations by tumor subtypes, particularly ER-positive versus ER-negative or triple negative (TN) disease[1-3]. However, prior GWAS have not simultaneously investigated multiple, correlated tumor markers to identify additional source(s) of etiologic heterogeneity. We performed a breast cancer GWAS using both standard analyses and a novel two-stage polytomous regression method that efficiently characterizes etiologic heterogeneity while accounting for tumor marker correlations and missing data[4].

The study populations and genotyping are described elsewhere[1,2,5,6] and in the **Online Methods**. Briefly, we analyzed data from 118,474 cases and 96,201 controls of European ancestry participating in 82 studies from the Breast Cancer Association Consortium (BCAC) and 9,414 affected and 9,494 unaffected BRCA*1* mutation carriers from 60 studies from the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) with genotyping data from one of two Illumina genome-wide custom arrays. In analyses of overall breast cancer, we also included summary level data from 11 other breast cancer GWAS (14,910 cases and 17,588 controls) without subtype information. Our study expands upon previous BCAC GWAS[1], with additional data on 10,407 cases and 7,815 controls, an approximate increase of 10% and 9%, respecitvely. (**Supplementary Tables 1-4**).

The statistical methods are further described in the **Online Methods** and in **Supplementary Figure 1**. To identify single nucleotide polymorphisms (SNPs) for overall breast cancer (invasive, *in situ* or unknown invasiveness) in BCAC, we used standard logistic regression to estimate odds ratios (OR) and 95% confidence-intervals

(CI) adjusting for country and principal components (PCs). iCOGS and OncoArray data were evaluated separately and results combined with those from the 11 other GWAS using fixed-effects meta-analysis.

To identify invasive breast cancer susceptibility SNPs displaying evidence of heterogeneity, we used a novel score-tests based on a two-stage polytomous model[4] that allows flexible, yet parsimonious, modelling of associations in the presence of underlying heterogeneity by ER, PR, HER2 and/or grade (**Online Methods, Supplementary Note**)[7]. The model handles missing tumor characteristic data by implementing an efficient Expectation-Maximization algorithm[4,8]. These analyses were restricted to BCAC controls and invasive cases (**Online Methods**). We fit an additional two-stage model to estimate case-control ORs and 95% CI between the SNPs and intrinsic-like subtypes defined by combinations of ER, PR, HER2 and grade (**Online Methods**): (1) luminal A-like, (2) luminal B/HER2-negative-like, (3) luminal B-like, (4) HER2-enriched-like and (5) TN or basal-like. We analyzed iCOGS and OncoArray data separately, adjusting for PCs and age, and meta-analyzed the results using a fixed-effects model. We evaluated the effect of country using a leave-one-out sensitivity analysis (**Online Methods**).

We used data from the *BRCA1* mutation carriers who are prone to develop TN disease[9], to estimate per-allele hazard ratios (HRs) within a retrospective cohort analysis framework. We assumed the estimated ORs for BCAC TN cases and the HRs estimated from CIMBA *BRCA1* carriers approximated the same underlying relative risk[9], and used a fixed-effect meta-analysis to combine these risk estimates (**Online Methods**). We used the two-stage polytomous model to test for heterogeneity in

associations for all newly identified SNPs across subtypes, globally and by tumor-specific markers (**Online Methods**).

Overall, we identified 32 novel independent susceptibility loci marked by SNPs with $P<5.0x10^{-8}$ (**Figure 1**, **Supplementary Table 5-7, Supplementary Figure 2-6**): 22 SNPs using standard logistic regression, eight SNPs using the two-stage polytomous model and three SNPs in the CIMBA/BCAC-TN meta-analysis (rs78378222 was also detected by the two-stage polytomous model in BCAC). Fourteen additional significant ($P<5.0x10^{-8}$) SNPs were excluded, 13 because they lacked evidence of association independent of previously reported susceptibility SNPs in conditional analyses ($P \geq 1.0x10^{-6}$; **Supplementary Table 8-10**), and one (chr22:40042814) for showing a high-degree of sensitivity to leave-one-out country analysis (**Supplementary Figure 7**). **Supplemental figures 8-9** show the associations between all 32 SNPs and the intrinsic-like subtypes.

Fifteen of the 32 SNPs showed evidence of heterogeneity (FDR<0.05) according to the global heterogeneity test (**Figure 2**, **Supplementary Table 11**). Nine of these were identified in analyses accounting for tumor marker heterogeneity. ER (7 SNPs) and grade (7 SNPs) most often contributed to observed heterogeneity (marker-specific $P<0.05$), followed by HER2 (4 SNPs) and PR (2 SNPs). rs17215231, identified in the CIMBA/BCAC-TN meta-analysis, was the only SNP found exclusively associated with TN disease (OR=0.85, 95%CI=0.81-0.89; $P=8.6x10^{-13}$). rs2464195, also identified in the CIMBA/BCAC-TN meta-analysis, was associated with both TN (OR=0.93, 95%CI=0.91-0.96; $P=2.5x10^{-8}$) and luminal B-like subtypes (OR=0.96, 95%CI=0.92-0.99; $P=0.02$; **Supplementary Table 7**, **Supplementary Figure 9**). This SNP is in LD ($r^2=0.62$) with

rs7953249, which is differentially associated with risk of subtypes of ovarian cancer[10]. Five of these heterogeneous SNPs showed associations with luminal and non-luminal subtypes in opposite directions (**Figure 3**). For example, four SNPs were associated in opposite directions with luminal A-like and TN subtypes (respectively, for rs78378222 OR=1.13, 95%CI=1.05-1.20 vs OR=0.67, 95%CI=0.57-0.80; for rs206435 OR=1.03, 95%CI=1.01-1.05 vs OR=0.95, 95%CI=0.92-0.98; for rs141526427 OR=0.96, 95%CI=0.94-0.98 vs OR=1.04, 95%CI=1.01-1.08; and for rs6065254 OR=0.96, 95%CI=0.94-0.97 vs OR=1.04, 95%CI=1.01-1.07). The specific tumor-marker heterogeneity test showed rs78378222 associated with ER (P_{ER}=7.0x10^{-6}) and HER2 (P_{HER2}=2.07x10^{-4}), rs206435 associated with ER (P_{ER}=2.8x10^{-3}) and grade (P_{grade}=2.8x10^{-4}) and rs141526427 (P_{ER}=1.3x10^{-3}) and rs6065254 (P_{ER}=4.3x10^{-3}) associated with ER. rs7924772 showed opposite associations between HER2-negative and HER2-positive subtypes (*e.g.*, OR=1.04, 95%CI=1.03-1.06 for luminal A-like disease and OR=0.95, 95%CI=0.92-0.99 for luminal B-like disease) and, consistent with these findings, was exclusively associated with HER2 (P_{HER2}=1.4x10^{-6}; **Figure 3**). Notably, rs78378222 located in the 3' UTR of *TP53* also showed opposite associations with high-grade serous cancers (OR=0.75, P=3.7x10^{-4}) and low-grade serous cancers (OR=1.58, P=1.5x10^{-4}; http://ocac.ccge.medschl.cam.ac.uk). Moreover, prior analyses did not find rs78378222 associated with risk of breast cancer, likely due to its opposite effects between subtypes[11].

We defined a set of candidate causal variants (CCVs; **Online Methods**) for each novel locus and investigated the CCVs in relation to previously-annotated enhancers in primary breast cells[12]. Based on combinations of H3K4me1 and H3K27ac histone

modification ChIP-seq signals, putative enhancers in basal cells (BC), luminal progenitor (LP) and mature luminal cells (LM) were characterized as “OFF,” “PRIMED”, and “ACTIVE” (**Online Methods**). We defined “ANYSWITCH” enhancers as those exhibiting different states between cell types. Among the five loci showing evidence of having associations in opposite directions between some subtypes, at least one CCV per locus overlapped an “ANYSWITCH” enhancer (**Figure 4**). For example, rs78378222 overlapped an ACTIVE enhancer in BC, PRIMED in LP and OFF in LM. In comparison, 63% of the loci with consistent direction of associations across subtypes overlapped with an “ANYSWITCH” enhancer (**Supplementary Table 12-13**). These results support the hypothesis that some variants may modulate enhancer activity in a cell-type specific manner and thus differentially influence the risk of developing different tumor subtypes.

We used INQUIST to intersect each of the CCVs with functional annotation data from public databases to identify potential target genes[1] (**Online Methods**, **Supplementary Table 14**). We predicted 179 unique target genes for 26 of the 32 independent signals. Twenty-three target genes in 14 regions were predicted with high confidence (designated “Level 1”), of which 22 target genes in 13 regions were predicted to be distally regulated. These targets include four genes predicted as INQUISIT targets in previous studies[13,14] *POLR3C*, *RNF115*, *SOX4* and *TBX3*, a known somatic breast cancer driver gene[15], and genes implicated by transcriptome-wide association studies (*LINC00886*[16] and *YBEY*[17]).

We used stratified LD-regression to investigate the genetic architecture of molecular subtypes by evaluating the genetic correlations[18,19] between subtypes and comparing enrichment of genomic features[20] between luminal A-like and TN subtypes

(**Online Methods**). All intrinsic-like subtypes were moderately- to highly-correlated, with luminal B/HER2-negative-like and TN subtypes (r=0.49, SE=0.06), and luminal A-like and TN (r=0.50, SE=0.04; **Figure 4**; **Supplementary Table 15**) having the lowest genetic correlations. Breast cancer in *BRCA1* mutation carriers and TN disease were highly genetically correlated (r=0.83, SE=0.08). To compare genomic enrichment, we first evaluated 53 annotations and found TN tumors were most enriched for "super-enhancers, extend500bp" (3.04-fold, $P = 3.3 \times 10^{-6}$), and "digital genomic footprint, extend500bp" (from DNase hypersensitive sites) (2.2-fold, $P = 4.0 \times 10^{-4}$) (**Supplementary Table 16, Supplementary Figure 10**). However, none of the 53 annotations significantly differed between luminal A-like and TN tumors. We also investigated cell-specific enrichment of four histone markers - H3K4me1, H3K3me3, H3K9ac and H3K27ac (**Online Methods**) - and found enrichment in both luminal-A and TN subtypes for gastrointestinal cell types and suppression of central nervous system cell types (**Supplementary Figure 11**).

The 32 identified SNPs explain approximately 1.2% of the two-fold familial relative risk for overall breast cancer. Collectively, the known and newly-identified common susceptibility SNPs explain approximately 18.3% of the familial relative risk. Moreover, we estimate that all common (MAF>0.01), reliably imputed variants on OncoArray can explain approximately 40.2% of the familial risk (**Online Methods**). The heritability explained by all identified susceptibility SNPs for the intrinsic-like subtypes ranged from 30.47% for HER2-enriched-like to 45.19% for luminal A-like, and for *BRCA1* carriers the explained heritability was 23.43% (**Table 1**). These analyses demonstrate the benefit of combining standard GWAS methods with methods

accounting for underlying tumor heterogeneity. Moreover, they may help clarify mechanisms predisposing to specific molecular subtypes, and provide precise risk estimates for molecular subtypes to inform the development of subtype-specific polygenic risk scores[21].

Online Methods

Study populations

The overall breast cancer analyses included women of European ancestry from 82 BCAC studies from over 20 countries, with genotyping data derived from two Illumina genome-wide custom arrays, the iCOGS and OncoArray (**Supplementary Table 1**). Most of the studies were case-control studies in the general population, or hospital setting, or nested within population-based cohorts, but a subset of studies oversampled cases with a family history of the disease. We included controls and cases of invasive breast cancer, carcinoma *in-situ*, and cases of unknown invasiveness. Information on clinicopathologic characteristics were collected by the individual studies and combined in a central database after quality control checks. We used BCAC database version 'freeze' 10 for these analyses. Among a subset of participants (n=16,766) that were genotyped on both the iCOGS and OncoArray arrays, we kept only the OncoArray data. One study (LMBC) contributing to the iCOGS dataset was excluded due to inflation of the test statistics that was not corrected by adjustment for the first ten PCs. We also excluded OncoArray data from Norway (the Norwegian Breast Cancer Study) because there were no controls available from Norway with OncoArray data. All participating studies were approved by their appropriate ethics or institutional review board and all participants provided informed consent. The total sample size for this analysis, including iCOGS, OncoArray and other GWAS data, comprised 133,384 cases and 113,789 controls.

In the GWAS analyses accounting for underlying heterogeneity according to ER, PR, HER2 and grade, we included genotyping data from 81 BCAC studies. These analyses were restricted to controls and cases of invasive breast cancer; we excluded cases of carcinoma *in-situ* and cases with missing information on invasiveness, as these cases would potentially bias the implicit "imputation" of tumor marker in the underlying EM algorithm (**Supplemental Table 2)**. We also excluded all studies from a specific country if there were no controls for that country, or if the tumor marker data were missing on two or more of the tumor marker subtypes (see footnote of **Supplemental Table 2** for further explanation of excluded studies). We did not include the summary results from the 14,910 cases and 17,588 controls from the 11 other GWAS in subtype analyses because these studies did not provide data on tumor characteristics. We also excluded invasive cases (n=293) and controls (n=4,285) with missing data on age at diagnosis or age at enrollment, information required by the EM algorithm to impute missing tumor characteristics. In total, the final sample for the two-stage polytomous logistic regression comprised 106,278 invasive cases and 91,477 controls.

Participants included from CIMBA were women of European ancestry, aged 18 years or older with a pathogenic *BRCA1* variant. Most participants were sampled through cancer genetics clinics. In some instances, multiple members of the same family were enrolled. OncoArray genotype data was available from 58 studies from 24 countries. Following quality control and removal of participants that overlapped with the BCAC OncoArray study, data were available on 15,566 *BRCA1* mutation carriers, of whom 7,784 were affected with breast cancer (**Supplementary Table 3**). We also

obtained iCOGS genotype data on 3,342 *BRCA1* mutation carriers (1,630 with breast cancer) from 54 studies through CIMBA. All *BRCA1* mutation carriers provided written informed consent and participated under ethically approved protocols.

Genotyping, quality control, and imputation

Details on genotype calling, quality control and imputation for the OncoArray, iCOGS, and GWAS are described elsewhere[1,2,5,6]. Genotyped or imputed SNPs marking each of the loci were determined using the iCOGS and the OncoArray genotyping arrays and imputation to the 1000 Genomes Project (Phase 3) reference panel. We included SNPs, from each component GWAS with an imputation quality score of >0.3. We restricted analysis to SNPs with a minor allele frequency >0.005 in the overall breast cancer analysis and >0.01 in the subtype analysis.

Known breast cancer susceptibility variants

Prior studies identified susceptibility SNPs from genome-wide analyses at a significance level $P< 5.0\times10^{-8}$ for all breast cancer types, ER-negative or ER-positive breast cancer, in *BRCA1* or *BRCA2* mutation carriers, or in meta-analyses of these[1,2]. We defined known breast cancer susceptibility variants as those variants that were identified or replicated in prior BCAC analyses[1,2]. We also excluded from consideration variants within 500kb of a previously published locus, since these regions have been subject to separate conditional analyses[14].

Standard analysis of BCAC data: Logistic regression analyses were conducted separately for the iCOGS and OncoArray datasets, adjusting for country and the array-specific first 10 PCs for ancestry informative SNPs. The methods for estimating PCs have been described elsewhere[1,2]. For the remaining GWAS, adjustment for inflation was done by adjusting for up to three PCs and using genomic control adjustment, as previously described[1]. We evaluated the associations between approximately 10.8 million SNPs with imputation quality scores (r^2) $\geqslant$0.3 and MAF >0.005. We excluded SNPs located within ±500 KB of, or in LD ($r^2 \geqslant 0.1$) with known susceptibility SNPs[22]. The association effect size estimates from these, and the previously derived estimates from the 11 other GWAS, were then combined using a fixed effects meta-analysis. Since individual level genotyping data were not available for some previous GWAS, we conservatively approximated the potential overlap between the GWAS and iCOGS and OncoArray datasets, based on the populations contributing to each GWAS (iCOGS/GWAS: 626 controls and 923 cases; OncoArray/GWAS: 20 controls and 990 cases). We then used these adjusted data to estimate the correlation in the effect size estimates, and incorporated these into the meta-analysis using the method of Lin and Sullivan[23].

Subtypes analysis of BCAC data: We described the two-stage polytomous logistic regression in more detail elsewhere[4,24] (**Supplementary Note**). In brief, this method allows for efficient testing of a SNP-disease association in the presence of tumor subtype heterogeneity defined by multiple tumor characteristics, while accounting for multiple testing and missing data on tumor characteristics. In the first stage, the model

uses a polytomous logistic regression to model case-control ORs between the SNPs and all possible subtypes that could be of interest, defined by the combination of the tumor markers. For example, in a model fit to evaluate heterogeneity according to ER, PR and HER2 positive/negative status, and grade of differentiation (low, intermediate and high grade), the first stage incorporates case-control ORs for 24 subtypes defined by the cross-classification of these factors. The second stage restructures the first-stage subtype-specific case-control ORs parameters into second-stage parameters through a decomposition procedure resulting in a second-stage baseline parameter that represents a case-control OR of a baseline cancer subtype, and case-case ORs parameters for each individual tumor characteristic. The second-stage case-case parameters can be used to perform heterogeneity tests with respect to each specific tumor marker while adjusting for the other tumor markers in the model. The two-stage model efficiently handles missing data by implementing an Expectation-Maximization algorithm[4,8] that essentially performs iterative "imputation" of the missing tumor characteristics conditional on available tumor characteristics and baseline covariates based on an underlying two-stage polytomous model.

To identify novel susceptibility loci, we used both a fixed-effect two-stage polytomous model and a mixed-effect two-stage polytomous model. The score-test we developed based on the mixed-effect model allows coefficients associated with individual tumor characteristics to enter as either fixed- or random-effect terms. Our previous analyses have shown that incorporation of random effect terms can improve power of the score-test by essentially reducing the effective degrees-of-freedom associated with fixed effects related to exploratory markers (*i.e.,* markers for which there

is little prior evidence to suggest that they are a source of heterogeneity)[25]. On the other hand, incorporation of fixed-effect terms can preserve distinct associations of known important tumor characteristics, such as ER. In the mixed-effect two-stage polytomous model, we therefore kept ER as a fixed effect, but modeled PR, HER2 and grade as random effects. We evaluated SNPs with MAF >0.01 (~9.7 million) and $r^2 \geqslant 0.3$, and excluded SNPs within $\pm$500 kb of, or in LD ($r^2 \geqslant 0.1$) with known susceptibility SNPs, including those identified in the standard analysis for overall breast cancer. We reported SNPs that passed the p-value threshold of $P < 5.0 \times 10^{-8}$ in either the fixed- or mixed-effects models.

We assessed the influence of country on signals identified by the two-stage models by performing a 'leave one out' sensitivity analyses in which we reevaluated novel signals after excluding data from each individual country. Data from the OncoArray and iCOGS arrays were analyzed separately and then meta-analyzed using fixed-effects meta-analysis.

Statistical analysis of CIMBA data: We tested for associations between SNPs and breast cancer risk for *BRCA1* mutation carriers using a score test statistic based on the retrospective likelihood of observing the SNP genotypes conditional on breast cancer phenotypes (breast cancer status and censoring time)[26]. Analyses were performed separately for iCOGS and OncoArray data. To allow for non-independence among related individuals, a kinship-adjusted test was used that accounted for familial correlations[27]. We stratified analyses by country of residence and, for countries where the strata were sufficiently large (United States and Canada), by Ashkenazi Jewish

ancestry. The results from the iCOGS and OncoArray data were then pooled using fixed-effects meta-analysis.

Meta-analysis of BCAC and CIMBA: We performed a fixed-effects meta-analysis of the results from BCAC TN cases and CIMBA *BRCA1* mutation carriers, using an inverse-variance fixed-effects approach implemented in METAL[28]. The estimates of association used were the logarithm of the per-allele hazard ratio estimate for association with breast cancer risk for *BRCA1* mutation carriers from CIMBA and the logarithm of the per-allele odds ratio estimate for association with risk of TN breast cancer based on BCAC data.

Conditional analyses: We performed two sets of conditional analyses. First, we investigated for evidence of multiple independent signals in identified loci by performing forward selection logistic regression, in which we adjusted the lead SNP and analyzed association for all remaining SNPs within ±500 kb of the lead SNPs, irrespective of LD. Second, we confirmed the independence of 20 SNPs that were located within ±2 MB of a known susceptibility region by conditioning the identified signals on the nearby known signal. Since these 20 SNPs are already genome-wide significant in the original GWAS scan and the conditional analyses restricted to local regions, we therefore used a significance threshold of $P<1\text{x}10^{-6}$ to control for type-one error[29].

Heterogeneity analysis of new association signals: We evaluated all novel signals for evidence of heterogeneity using two-stage polytomous model. We first performed a global test for heterogeneity under the mixed-effect model test to identify SNPs showing evidence of heterogeneity with respect to any of the underlying tumor markers, ER, PR, HER2 and/or grade. We accounted for multiple testing of the global

heterogeneity test using a FDR <0.05 under the Benjamini-Hochberg procedure[30]. Among the SNPs with observed heterogeneity, we then further used a fixed-effect two-stage model to evaluate influence of specific tumor characteristic(s) driving observed heterogeneity, adjusted for the other markers in the model. We also fit a separate fixed-effect two-stage models to estimate case-control ORs and 95% confidence intervals (CI) for five surrogate intrinsic-like subtypes defined by combinations of ER, PR, HER2 and grade[31]: (1) luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); (2) luminal B,HER2-negative-like (ER+ and/or PR+, HER2-, grade 3); (3) luminal B-like (ER+ and/or PR+, HER2+); (4) HER2-enriched-like (ER- and PR-, HER2+), and (5) TN (ER-, PR-, HER2-).

Effective sample size of cases of two-stage polytomous model

The two-stage polytomous model implements the EM algorithm to impute missing tumor characteristics; therefore, the effective sample size of cases is not equivalent to the actual number of cases with available tumor characteristic data. We estimated the effective sample sizes to help demonstrate the benefit of using the EM algorithm to impute missing tumor characteristics and to aid comparability with previous studies (**Supplementary Table 4**). To estimate the effective sample size, suppose we have a complete dataset with no missing tumor characteristics, the sample size is $\mathrm{n_k}$ for the kth subtype and $\mathrm{n_0}$ for the control. If we fit a two-stage polytomous model for the jth SNP, the corresponding log odds ratio for kth subtype is $\hat{\beta}_{jk}$ and the standard error is $\mathrm{s_{jk}}$. Then, approximately:

$$var\left(\hat{\beta}_{jk}|p_j\right) = \frac{n_0 + n_k}{2 * p_j\left(1 - p_j\right)(n_0 n_k)},$$

where p_j is the MAF of the jth SNP. Now we consider fitting a two-stage polytomous model with missing tumor characteristics. Given the standard error s_{jk} of the log odds ratio and the control sample size, we have the estimate of effective number of cases as,

$$\hat{n}_k = \left(\frac{1}{n_0} - 2s_{jk}^2 p_j(1-p_j)\right)^{-1}.$$

We used the median estimates of effective sample size of cases for all SNPs as the final estimate.

Candidate causal variants

We defined credible sets of candidate causal variants (CCVs) as variants located within ±500kb of the lead SNPs in each novel region and with *P* values within 100-fold of magnitude of the lead SNPs. This is approximately equivalent to selecting variants whose posterior probability of causality is within two orders of magnitude of the most significant SNP[32,33]. This approach was applied for detecting a set of potentially causal variants for all 32 identified SNPs. For the novel SNPs located within ±2Mb of the known signals, we used the conditional *P* values to adjust for the known signals' associations.

eQTL Analysis

Data from breast cancer tumors and adjacent normal breast tissue were accessed from The Cancer Genome Atlas (TCGA)[34]. Germline SNP genotypes (Affymetrix 6.0 arrays) were processed and imputed to the 1000 Genomes reference panel (October 2014) and European ancestry ascertained as previously described[1]. Tumor tissue copy number was estimated from the Affymetrix 6.0 and called using the

GISTIC2 algorithm[35]. Complete genotype, RNA-seq and copy number data were available for 679 genetically European patients (78 with adjacent normal tissue). Further, RNA-seq for normal breast tissue and imputed germline genotype data were available from 80 females from the GTEx Consortium[36]. Genes with a median expression level of 0 RPKM across samples were removed, and RPKM values of each gene were log2 transformed. Expression values of samples were quantile normalized. Genetic variants were evaluated for association with the expression of genes located within ±2Mb of the lead variant at each risk region using linear regression models, adjusting for ESR1 expression. Tumor tissue was also adjusted for copy number variation, as previously described[37]. eQTL analyses were performed using the MatrixEQTL program in R[38].

INQUISIT target gene analysis

Logic underlying INQUISIT predictions: Details of the INQUISIT pipeline have been previously described[1]. Briefly, genes were evaluated as potential targets of candidate causal variants through effects on: (1) distal gene regulation, (2) proximal regulation, or (3) a gene's coding sequence. We intersected CCV positions with multiple sources of genomic information, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[39] in MCF7 cells, and genome-wide chromosome conformation capture (Hi-C) in HMECs[40]. We used breast cell line computational enhancer–promoter correlations (PreSTIGE[41], IM-PET[42], FANTOM5[43]) breast cell super-enhancer[44], breast tissue-specific expression variants (eQTL) from multiple independent studies (TCGA (normal breast and breast tumor) and GTEx breast, **See eQTL Methods**), transcription

factor and histone modification chromatin immunoprecipitation followed by sequencing (ChIP-seq) from the ENCODE and Roadmap Epigenomics Projects together with the genomic features found to be significantly enriched for all known breast cancer CCVs[14], gene expression RNA-seq from several breast cancer lines and normal samples (ENCODE) and topologically associated domain (TAD) boundaries from T47D cells (ENCODE[45]). To assess the impact of intragenic variants, we evaluated their potential to alter primary protein coding sequence and splicing using Ensembl Variant Effect Predictor[46] using MaxEntScan and dbscSNV modules for splicing alterations based on "ada" and "rf" scores. Nonsense and missense changes were assessed with the REVEL ensemble algorithm, with CCVs displaying REVEL scores > 0.5 deemed deleterious.

Scoring hierarchy: Each target gene prediction category (distal, promoter or coding) was scored according to different criteria. Genes predicted to be distally-regulated targets of CCVs were awarded two points based on physical links (for example ChIA-PET), and one point for computational prediction methods, or eQTL associations. All CCVs were considered as potentially involved in distal regulation and all CCVs (including coding SNPs) were scored in this category. Intersection of a putative distal enhancer with genomic features found to be significantly enriched[20] were further upweighted with an additional point. In the case of multiple, independent interactions, an additional point was awarded. CCVs in gene proximal regulatory regions were intersected with histone ChIP-Seq peaks characteristic of promoters and assigned to the overlapping transcription start sites (defined as -1.0 kb - +0.1 kb). Further points were awarded to such genes if there was evidence for an eQTL association, while a lack of expression resulted in down-weighting as potential targets.

Potential coding changes including missense, nonsense and predicted splicing alterations resulted in addition of one point to the encoded gene for each type of change, while lack of expression reduced the score. We added an additional point for predicted target genes that were also breast cancer drivers (278 genes[1,20]). For each category, scores potentially ranged from 0-8 (distal); 0-4 (promoter) or 0-3 (coding). We converted these scores into 'confidence levels': Level 1 (highest confidence) when distal score >4, promoter score ≥3 or coding score >1; Level 2 when distal score ≤4 and ≥1, promoter score=1 or=2, coding score=1; and Level 3 when distal score <1 and >0, promoter score <1 and >0, and coding <1 and >0. For genes with multiple scores (for example, predicted as targets from multiple independent risk signals or predicted to be impacted in several categories), we recorded the highest score.

Enhancer states analysis in breast sub-populations

We obtained enhancer maps for three enriched primary breast sub-populations (basal, luminal progenitor, and mature luminal) from Pellacani et al.[12]. Enhancer annotations were defined as ACTIVE, PRIMED, or OFF based on a combination of H3K27ac and H3K4me1 histone modification ChIP-seq signals using FPKM thresholds as previously described[12]. Briefly, genomic regions containing high H3K4me1 signal observed in any cell type were used to define the superset of breast regulatory elements. Sub-population cell type-specific H3K27ac signal (which is characteristic of active elements) within these elements was used as a measure of overall regulatory activity, where "ACTIVE" sites were characterized by H3K4me1-high, H3K27ac-high; "PRIMED" by H3K4me1-high, H3K27ac-low; and "OFF" by H3K4me1-low, H3K27ac-low. This enabled annotation of each enhancer element as either “OFF”, “PRIMED” or

“ACTIVE” in all cell types. We then defined enhancers which exhibit differing states between at least one cell type as "ANYSWITCH" enhancers.

Genetic correlation analyses

We used LD score regression[18-20] to assess the heritability due to susceptibility SNPs and estimated the genetic correlation between five intrinsic-like breast cancer subtypes. The analysis used the summary statistics based on the meta-analysis of the OncoArray, and iCOGS, and CIMBA meta-analysis. The genetic correlation[18] analysis was restricted to the ~1 million SNPs included in HapMap 3. Since two-stage polytomous models integrated an imputation algorithm for missing tumor characteristic data, we modified the LD score regression to generate the effective sample size for each SNP (**Supplementary Note**).

Global genomic enrichment analyses

We performed stratified LD score regression analyses[20] as previously described[1] for two major intrinsic-like subtypes, luminal A-like and TN, using the summary statistics from the meta-analyses of OncoArray, iCOGs, and CIMBA. The analysis included all SNPs in the 1000 Genome Project Phase 1v3 release with MAF>1% and imputation quality score R2>0.3 in the OncoArray data. We first fit a model that included 24 non-cell-type-specific, publicly available annotations as well as 24 additional annotations that included a 500-bp window around each of the 24 main annotations. We also included 100-bp windows around ChIP–seq peaks and one annotation containing all SNPs, leading to a total of 53 overlapping annotations. In addition to the baseline model using

24 main annotations, we also performed cell-type-specific analyses using annotations of the four histone marks (H3K4me1, H3K4me3, H3K9ac and H3K27ac). Each cell-type-specific annotation corresponds to a histone mark in a single cell type (for example, H3K27ac in adipose nuclei tissues)[20]. There was a total of 220 such annotations. We further subdivided these 220 cell-type-specific annotations into 10 categories by aggregating the cell-type-specific annotations within each group (for example, SNPs related with any of the four histone modifications in any hematopoietic and immune cells were considered as one category). To estimate the enrichment of each marker, we ran 220 LD score regressions after adding each different histone mark to the baseline model. We used a Wald test to evaluate the differences in the functional enrichment between the luminal A-like and TN subtypes, using the regression coefficients and standard error based on the models above. After Bonferroni correction none of the differences were significant. Notably, the Wald test assumes that the enrichment estimates of luminal A-like and TN subtypes were independent, but this assumption was violated by the sharing of controls between the subtypes. Under this scenario, our Wald test statistics were less conservative than had we adjusted for the correlation between estimates. However, given the lack of significant differences observed between luminal A-like and TN subtypes we had no concern about a type one error.

Contribution of identified variants to the familial relative risk of breast cancer

We define the familial relative risk as λ. Under a log-additive model, we define the heritability as σ^2, and the relationship between λ and σ^2 as $\sigma^2 = 2 * \log(\lambda)$ [47]. Under

the log-additive model, the frailty-scale heritability that is explained by the identified variants can be estimated by:

$$\sum_{i=1}^{n} 2p_i(1-p_i)(\hat{\beta}_{\mathrm{i}}^2 - \tau_i^2),$$

where n is the total number of identified SNPs, $\mathrm{p_i}$ is the MAF for variant I, $\hat{\beta}_{\mathrm{i}}$ is the log odds ratio estimate for the variant I, and τ_{i} is the standard error of $\hat{\beta}_{\mathrm{i}}$. The corresponding frailty-scale heritability for all variants is and $\sigma^2 = 2 * \log(\lambda)$, where λ is the familial relative risk to first degree relatives of affected individuals, assuming a polygenic log-additive model that explains all the familial aggregation of the disease[47]. We assumed $\lambda = 2$ as the overall familial relative risk of breast cancer, so the proportion of the familial relative risk explained by the identified SNPs is $\sum_{i=1}^{n} p_i(1-p_i)(\hat{\beta}_i^2 - \tau_i^2)/\log(2)$. To obtain the heritability explained by all of the GWAS variants, we estimated the heritability (σ^2_{GWAS}) using the full set of summary statistics using LD score regression as previously described1. σ^2_{GWAS} is characterized by population variance of the underlying true polygenetic risk scores as $h^2 = Var(\sum_{m=1}^{M} \beta_m G_m)$, where β_m is the true log odds ratio for the $m\mathrm{th}$ SNP. The proportion of the familial relative risk explained by GWAS variants is $\sigma^2_{GWAS}\ /[2 * \log(2)]$. Thus, the proportion of heritability explained by identified variants relative to all imputable SNPs is:

$$\sum_{i=1}^{n} 2p_i(1-p_i)(\hat{\beta}_{\mathrm{i}}^2 - \tau_i^2)\ /\sigma^2_{GWAS}.$$

References

1. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94 (2017).
2. Milne, R.L. *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* **49**, 1767-1778 (2017).
3. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45**, 392-8, 398e1-2 (2013).
4. Zhang, H. *et al.* A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *bioRxiv*, 446039 (2018).
5. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-61, 361e1-2 (2013).
6. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373-80 (2015).
7. Zhang, B. *et al.* Height and Breast Cancer Risk: Evidence From Prospective Studies and Mendelian Randomization. *J Natl Cancer Inst* **107**(2015).
8. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological* **39**, 1-38 (1977).
9. Spurdle, A.B. *et al.* Refined histopathological predictors of BRCA1 and BRCA2 mutation status: a large-scale analysis of breast cancer characteristics from the BCAC, CIMBA, and ENIGMA consortia. *Breast Cancer Res* **16**, 3419 (2014).
10. Phelan, C.M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **49**, 680-691 (2017).
11. Stacey, S.N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat Genet* **43**, 1098-103 (2011).
12. Pellacani, D. *et al.* Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Rep* **17**, 2060-2074 (2016).
13. Beesley, J. *et al.* Chromatin interactome mapping at 139 independent breast cancer risk signals. *bioRxiv*, 520916 (2019).
14. Fachal, L. *et al.* Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. *bioRxiv*, 521054 (2019).
15. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).
16. Ferreira, M.A. *et al.* Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat Commun* **10**, 1741 (2019).
17. Wu, L. *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* **50**, 968-978 (2018).
18. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
19. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
20. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
21. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21-34 (2019).
22. Ahearn, T.U. *et al.* Common breast cancer risk loci predispose to distinct tumor subtypes. *bioRxiv*, 733402 (2019).

23. Lin, D.Y. & Sullivan, P.F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* **85**, 862-72 (2009).
24. Chatterjee, N. A Two-Stage Regression Model for Epidemiological Studies with Multivariate Disease Classification Data. *Journal of the American Statistical Association* **99**, 127-138 (2004).
25. Zhang, H. *et al.* A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *bioRxiv*, 446039 (2018).
26. Barnes, D.R. *et al.* Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations. *Genet Epidemiol* **36**, 274-91 (2012).
27. Antoniou, A.C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**, 885-92 (2010).
28. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
29. Hendricks, A.E., Dupuis, J., Logue, M.W., Myers, R.H. & Lunetta, K.L. Correction for multiple testing in a gene region. *Eur J Hum Genet* **22**, 414-8 (2014).
30. 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)* **57**, 289-300 (1995).
31. Curigliano, G. *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* **28**, 1700-1712 (2017).
32. Udler, M.S., Tyrer, J. & Easton, D.F. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet Epidemiol* **34**, 463-8 (2010).
33. Wellcome Trust Case Control, C. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).
34. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
35. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
36. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
37. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
38. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).
39. Fullwood, M.J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).
40. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
41. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* **24**, 1-13 (2014).
42. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* **111**, E2191-9 (2014).
43. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).
44. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).

45. Dixon, J.R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet* **50**, 1388-1398 (2018).
46. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
47. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33-6 (2002).