Association of Breast Cancer Risk with Genetic Variants Showing Differential Allelic Expression: Identification of a Novel Breast Cancer Susceptibility Locus at 4q21.

Yosr Hamdi[1¶], Penny Soucy[1¶], Véronique Adoue[2,3,4], Kyriaki Michailidou[5,6], Sander Canisius[7], Audrey Lemaçon[8], Arnaud Droit[8], Irene L Andrulis[9,10], Hoda Anton-Culver[11], Volker Arndt[12], Caroline Baynes[13], Carl Blomqvist[14], Natalia V. Bogdanova[15,16], Stig E. Bojesen[17,18,19], Manjeet K. Bolla[5], Bernardo Bonanni[20], Anne-Lise Borresen-Dale[21,22], Judith S. Brand[23], Hiltrud Brauch[24,25,26], Hermann Brenner[12,26,27], Annegien Broeks[7], Barbara Burwinkel[28,29], Jenny Chang-Claude[30,31], NBCS Collaborators[32-35,21,36-41,42], Fergus J. Couch[43], Angela Cox[44], Simon S. Cross[45], Kamila Czene[23], Hatef Darabi[23], Joe Dennis[5], Peter Devilee[46,47], Thilo Dörk[16], Isabel Dos-Santos-Silva[48], Mikael Eriksson[23], Peter A. Fasching[49,50], Jonine Figueroa[51,52], Henrik Flyger[53], Montserrat García-Closas[52], Graham G. Giles[54,55], Mark S. Goldberg[56,57], Anna González-Neira[58], Grethe Grenaker-Alnæs[21], Pascal Guénel[59], Lothar Haeberle[49], Christopher A. Haiman[60], Ute Hamann[61], Emily Hallberg[62], Maartje J. Hooning[63], John L. Hopper[55], Anna Jakubowska[64], Michael Jones[65], Maria Kabisch[61], Vesa Kataja[66,67], Diether Lambrechts[68,69], Loic Le Marchand[70], Annika Lindblom[71], Jan Lubinski[64], Arto Mannermaa[66,72,73], Mel Maranian[13], Sara Margolin[74], Frederik Marme[75,28], Roger L. Milne[54,55], Susan L. Neuhausen[76], Heli Nevanlinna[77], Patrick Neven[78], Curtis Olswold[62], Julian Peto[48], Dijana Plaseska-Karanfilska[79], Katri Pylkäs[80,81], Paolo Radice[82], Anja Rudolph[30], Elinor J. Sawyer[83], Marjanka K. Schmidt[7], Xiao-Ou Shu[84], Melissa C. Southey[85], Anthony Swerdlow[86], Rob A. E. M. Tollenaar[87], Ian Tomlinson[88], Diana Torres[89,61], Thérèse Truong[59], Celine Vachon[62], Ans M. W. Van Den Ouweland[90], Qin Wang[5], Robert Winqvist[80,81], kConFab/AOCS Investigators[91], Wei Zheng[84], Javier Benitez[58,92], Georgia Chenevix-Trench[93], Alison M. Dunning[13], Paul D. P. Pharoah[5,13], Vessela Kristensen[21,22,94], Per Hall[23], Douglas F. Easton[5,13], Tomi Pastinen[95,96], Silje Nord[97,98], Jacques Simard[1*].

[1] Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Quebec, G1V 4G2, Canada.
[2] Institut National de la Santé et de la Recherche Médicale U1043, Toulouse, 31024 France
[3] Centre National de la Recherche Scientifique U5282, Toulouse, 31400, France
[4] Université de Toulouse, Université Paul Sabatier, Centre de Physiopathologie de Toulouse Purpan, Toulouse, 31024, France
[5] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK,
[6] Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, 1683, Cyprus
[7] Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital, Amsterdam, 1066, The Netherlands,
[8] Centre de Recherche du CHU de Québec – Université Laval, Faculté de Médecine, Département de Médecine Moléculaire, Université Laval, Quebec, G1V 4G2, Canada
[9] Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, M5G 1X5, Canada
[10] Department of Molecular Genetics, University of Toronto, Toronto, M5S 1A8, Canada
[11] Department of Epidemiology, University of California Irvine, Irvine, 92697, CA, USA

[12] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, 69121, Germany
[13] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, CB1 8RN, UK
[14] Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland
[15] Department of Radiation Oncology, Hannover Medical School, Hannover, 00014, Germany
[16] Gynaecology Research Unit, Hannover Medical School, Hannover, 30625, Germany
[17] Copenhagen General Population Study, Herlevand Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark
[18] Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark
[19] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2200, Denmark
[20] Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, Milan, 2014, Italy
[21] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, 0372, Norway
[22] K.G. Jebsen Center for Breast Cancer Research, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, 0450, Norway
[23] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, 104 35, Sweden
[24] Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany,
[25] University of Tübingen, Tübingen, 70376, Germany
[26] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany
[27] Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, 69120, Germany
[28] Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, 00014, Germany
[29] Molecular Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, 69121, Germany
[30] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, 69121, Germany
[31] University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, 20246, Germany
[32] Department of Oncology, Haukeland University Hospital, Bergen, 5021, Norway
[33] Section of Oncology, Institute of Medicine, University of Bergen, Bergen, 5020, Norway
[34] Department of Pathology, Akershus University Hospital, Lørenskog, 1478, Norway
[35] Department of Breast-Endocrine Surgery, Akershus University Hospital, Lørenskog, 1478, Norway
[36] Department of Breast and Endocrine Surgery, Oslo University Hospital, Ullevål, Oslo, 1478, Norway
[37] Department of Research, Vestre Viken, Drammen, 3004, Norway
[38] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, 0450, Norway
[39] National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital Radiumhospitalet, Oslo, NO-0310, Norway

[40] Department of Oncology, Oslo University Hospital Radiumhospitalet, Oslo, NO-0424, Norway
[41] Department of Radiology and Nuclear Medicine, Oslo University Hospital Radiumhospitalet, Oslo, 0372, Norway
[42] Oslo University Hospital, Oslo, 1478, Norway
[43] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, 55905, USA
[44] Sheffield Cancer Research, Department of Oncology and Metabolism, University of Sheffield, Sheffield, S10 2TN, UK
[45] Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, S10 2TN, UK
[46] Department of Pathology, Leiden University Medical Center, Leiden, 2333, The Netherlands
[47] Department of Human Genetics, Leiden University Medical Center, Leiden, 2333, The Netherlands
[48] Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
[49] Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, 40225, Germany
[50] David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, 90095, USA
[51] Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, EH16 4TJ, UK
[52] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 9609, USA
[53] Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark
[54] Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, 3004, Australia
[55] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global health, The University of Melbourne, Melbourne, 3010, Australia
[56] Department of Medicine, McGill University, Montreal, H3G 2M1, Canada
[57] Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montreal, H3A 1A2, Canada
[58] Human Cancer Genetics Program, Spanish National Cancer Research Centre, Madrid, E-28029, Spain
[59] Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, VilleJuif, 91405, France
[60] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, 90089, USA
[61] Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, 69121, Germany
[62] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55905, USA
[63] Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, 3318, The Netherlands
[64] Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, 70-204 Poland

[65] Division of Genetics and Epidemiology, the Institute of Cancer Research, London, SM2 5NG, UK
[66] Cancer Center of Eastern Finland, University of Eastern Finland, Kuopio, 80130, Finland
[67] Central Finland Hospital District, Jyväskylä Central Hospital, Jyväskylä, 40620, Finland
[68] Vesalius Research Center, Leuven, 3000, Belgium
[69] Laboratory for Translational Genetics, Department of Oncology, University of Leuven, Leuven, 3000, Belgium
[70] University of Hawaii Cancer Center, Honolulu, HI, 96813, USA
[71] Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, 104 35, Sweden
[72] Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, 80130, Finland
[73] Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, 70210, Finland
[74] Department of Oncology - Pathology, Karolinska Institutet, Stockholm, 104 35, Sweden
[75] National Center for Tumor Diseases, University of Heidelberg, Heidelberg, 69120, Germany
[76] Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA
[77] Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, 00014, Finland
[78] Multidisciplinary Breast Center, Department of Oncology, University Hospitals Leuven, Leuven, 3000, Belgium
[79] Research Center for Genetic Engineering and Biotechnology "Georgi D. Efremov", Macedonian Academy of Sciences and Arts, 1000, Skopje, Republic of Macedonia
[80] Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Aapistie 5 A, 90220, Oulu, Finland
[81] Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Aapistie 5 A, 90220, Oulu, Finland
[82] Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS (Istituto Di Ricovero e Cura a Carattere, Scientifico) Istituto Nazionale Tumori (INT), Milan, 20133, Italy.
[83] Research Oncology, Guy's Hospital, King's College London, London, SE1 9RT, UK
[84] Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, 37232, USA
[85] Department of Pathology, The University of Melbourne, Melbourne, 3065, Australia
[86] Division of Genetics and Epidemiology & Division of Breast Cancer Research, The Institute of Cancer Research, London, SM2 5NG, UK
[87] Department of Surgery, Leiden University Medical Center, Leiden, 2333, The Netherlands
[88] Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, OX3 7BN, UK
[89] Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, 40 - 62, Colombia
[90] Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, 3015, The Netherlands
[91] Peter MacCallum Cancer Center, the University of Melbourne, Melbourne, 3002, Australia
[92] Centro de Investigación en Red de Enfermedades Raras, Valencia, 28029, Spain
[93] Department of Genetics, QIMR Berghofer Medical Research Institute, Brisbane, 4006, Australia

[94] Department of Clinical Molecular Biology, Oslo University Hospital, University of Oslo, Oslo, 1478, Norway
[95] Department of Human Genetics, McGill University, Montreal, Quebec, H3A 1B1, Canada
[96] McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, H3A 0G1, Canada
[97] Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Radiumhospitalet, Ullernchausseen, Oslo, 0372, Norway
[98] K.G. Jebsen Center for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Kirkeveien, Oslo, 0450, Norway

[¶] These authors contributed equally to this work

[*]Corresponding author:

Jacques Simard
Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Quebec, Canada
Phone: 418-654-2264
Fax: 418-654-2278
Email: Jacques.Simard@crchudequebec.ulaval.ca

Keywords: Breast Cancer, Genetic Susceptibility, Association Studies, Differential Allelic Expression, cis-regulatory variants, low-penetrance alleles

Note: This manuscript includes 6 figures and 1 table, as well as 3 supplementary figures and 6 supplementary tables

Disclosure of potential conflicts of interest
We do not expect any conflict of interests to disclose. If the manuscript moves to the revision stage, we will promptly coordinate the completion of the Conflict of Interest forms for all co-authors and provide these to the editorial office.

Abstract

There are significant inter-individual differences in the levels of gene expression. Through modulation of gene expression, *cis*-acting variants represent an important source of phenotypic variation. Consequently, c*is*-regulatory SNPs associated with differential allelic expression are functional candidates for further investigation as disease-causing variants. To investigate whether common variants associated with differential allelic expression were involved in breast cancer susceptibility, a list of genes was established on the basis of their involvement in cancer related pathways and/or mechanisms. Thereafter, using data from a genome-wide map of allelic expression associated SNPs, 313 genetic variants were selected and their association with breast cancer risk was evaluated then in 46,451 breast cancer cases and 42,599 controls of European ancestry ascertained from 41 studies participating in the Breast Cancer Association Consortium. The associations were evaluated with overall breast cancer risk and with estrogen receptor negative and positive disease. One novel breast cancer susceptibility locus on 4q21 (rs11099601) was identified (OR=1.05, P=5.6x10^{-6}). rs11099601 lies in a 135 kb linkage disequilibrium block containing several genes, including, *HELQ*, encoding the protein HEL308 a DNA dependant ATPase and DNA Helicase involved in DNA repair, *MRPS18C* encoding the Mitochondrial Ribosomal Protein S18C and *FAM175A (ABRAXAS)*, encoding a *BRCA1* BRCT domain-interacting protein involved in DNA damage response and double-strand break (DSB) repair. Expression QTL analysis in breast cancer tissue showed rs11099601 to be associated with *HELQ* (P=8.28x10^{-14}), *MRPS18C* (P=1.94x10^{-27}) and *FAM175A* (P=3.83x10^{-3}) and explaining respectively about 20%, 14% and 1% of the variance in expression of these genes in breast carcinomas.

Introduction

Breast cancer is a complex disease with a strong heritable component. Great efforts have been made during the last decades to elucidate the underlying etiology of this disease. Three classes of breast cancer susceptibility alleles with different levels of risk and prevalence in the population are now recognized. High-risk alleles such as *BRCA1* [1, 2], *BRCA2* [3, 4] and *TP53* [5] explain approximately 20% of the inherited susceptibility, intermediate-risk alleles in DNA-repair genes increase this proportion by ~5% [6-18], and common lower-risk alleles, of which approximately 100 have been identified to date through genome-wide association studies (GWAS), replication and custom genotyping efforts, explain approximately 16% of the risk [19-41]. Recent evidence suggests that a substantial fraction of the residual aggregation could be explicable by other common variants not yet identified [35, 40].

Global analysis of genome-wide association study (GWAS) data has shown that the large majority of common variants associated with susceptibility to cancer lie in non-coding regions, and are presumed to mediate risk through regulation of gene expression [42, 43]. Indeed, variations in gene expression occur commonly in the human genome, playing a key role in human phenotypic variability [44-46]. Studies of allelic imbalances in expression indicate that allele-specific differences among transcripts within an individual can affect up to 30% of loci and, at the population level, ~30% of expressed genes show evidence of *cis*-regulation by common polymorphic alleles [47]. Recent evidence has also suggested that differences in gene expression play a critical role in the underlying phenotypic variation associated with many complex genetic diseases [48]. A recent report performed expression quantitative trait loci (*cis*-eQTL) analyses for mRNA expression in five tumor types (breast, colon, kidney, lung and prostate) and tested 149 known cancer risk loci for eQTL effects [49]. They observed that 42 of

these risk loci were significantly associated with eQTLs in at least one gene within 500 kb, eight of which were breast cancer risk loci [49]. Furthermore, a recent study has shown that close to half of the known risk alleles for estrogen receptor (ER)-positive breast cancer are eQTLs acting upon major determinants of gene expression in tumors [50]. These results suggest that additional cancer susceptibility loci may be identified through studying genetic variants affecting regulation of gene expression.

In the current study, we performed a breast cancer association study of 313 genetic variants showing evidence of association with differential allelic expression (DAE) selected from 175 genes involved in cancer etiology. These included genes involved in DNA repair (homologous recombination (HR) and DNA interstrand crosslink (ICL) repair), interacting and/or modulating BRCA1 and BRCA2 cellular functions, cell cycle control, centrosome amplification and AURKA interactions, apoptosis, ubiquitination, known tumor suppressors and mitotic and other kinases, as well as sex steroid action and mammographic density. We used genotype data derived from the iCOGS (Collaborative Oncological Gene-environment Study) custom array [35] to investigate the role of these variants on breast cancer risk.

Results

Overall and subtype-specific breast cancer risk association analyses

For the one hundred seventy-five selected genes involved in cancer-related pathways, we identified a set of 355 genetic variants showing evidence of association with DAE (see S1 Table for complete list of genes and SNPs). Of the 355 SNPs originally selected, 313 (representing 227 independent SNPs with pairwise $r^2<0.1$) were successfully genotyped. Thirty-two variants were excluded because of low Illumina design scores, and eleven SNPs were excluded because of low call-rates and/or evidence of deviation from Hardy Weinberg Equilibrium (P-value$<10^{-7}$), respectively. Eighty-two SNPs were originally submitted to be included on the iCOGS array but were replaced with surrogates in the final design of the array. Association results with breast cancer risk for all 313 SNPs are presented in S2 Table.

Thirteen SNPs from ten different loci were associated with overall breast cancer risk ($P<10^{-2}$) (Table 1). Of these, three SNPs, namely rs11099601, rs656040 and rs738200, had associations with an increased overall risk of breast cancer that reached $P<10^{-4}$ (approximate significance cut-off after Bonferroni correction, given 313 tests). No significant evidence of heterogeneity was observed among odds ratios (ORs) for these SNPs among studies (I^2 and P-values are given in S1 Fig.). The minor alleles of rs11099601 at 4q21 (OR=1.05, $P=5.6\text{x}10^{-6}$), rs656040 at 11q13 (OR=1.05, $P=1.52\ \text{x}10^{-5}$), and rs738200 at 22q12.1 (OR=1.09, $P=5.32\text{x}10^{-5}$) were associated with increased overall risk of the disease. rs11099601 was associated with both ER-positive ($P=5.22\text{x}10^{-6}$) and ER-negative ($P=4.08\text{x}10^{-4}$) breast cancer risk (P for difference 0.93) while rs656040 and rs738200 appeared primarily associated with ER-positive disease ($P=5.96\text{x}10^{-5}$ and $P=7.21\text{x}10^{-6}$, respectively), although the difference between ER-positive and ER-negative disease

was not statistically significant for these two latter SNPs (*P* for difference 0.096 and 0.242, respectively). Of these three SNPs, only variant rs110099601 represents a novel low penetrance breast cancer susceptibility locus. The two other variants, (rs656040 at 11q13 and rs738200 at 22q12.1) which were not known to be associated with breast cancer risk at the time the current study was designed, were identified through the main analyses of the iCOGS array. rs656040 is located on 11q13 in the 3'-UTR region of the *SNX32* gene, approximately 6.8Kb upstream of *MUS81,* and is associated with differential allelic expression of this latter gene (S2 Fig.). rs656040 is partially correlated with rs3903072 (r^2=0.38), which was previously identified as associated with breast cancer risk at $P<10^{-8}$ in the combined GWAS and iCOGS analysis reported in Michailidou et al. [35]. Similarly, variant rs738200, located on locus 22q12 in the tetratricopeptide repeat domain 28 gene (*TTC28*), falls within a 610 kb interval (Build 37 coordinates chr22: 28,314,612–28,928,858) on chromosome 22 recently shown to be associated with breast cancer risk (smallest $P=8.2\times10^{-22}$, for rs62237573). This interval lies approximately 100 kb centromeric to *CHEK2*, and further analysis showed that the associated SNPs were correlated with the deleterious *CHEK2* variant c.1100delC and adjustment for this variant suggested the signal is driven by *CHEK2* c.1100delC [40]. rs738200 was genotyped as a surrogate to our originally selected SNP for this locus (rs9620797), and therefore no allelic expression data were available for this SNP.

All variants associated with overall breast cancer risk with $P<10^{-2}$ included in Table 1 were also evaluated for association with breast cancer risk in *BRCA1* and *BRCA2* mutation carriers within the Consortium of Investigators of Modifiers of *BRCA1* and *BRCA2* (CIMBA) in a total of 15 252 *BRCA1* and 8 211 *BRCA2* carriers. However, none of the SNPs showed associations with breast cancer risk, including rs11099601, which had a *P*-value of 0.89 and 0.78 in *BRCA1* and *BRCA2* carriers respectively.

rs11099601 lies on 4q21 in a region containing numerous genes including *FAM175A* (*ABRAXAS)*, *HELQ* and *MRPS18C*. It was selected on the basis of its association with differential allelic expression in *FAM175A* (see S2 Fig.). In order to further map the novel association at this locus, we imputed genotype data for 2,456 common variants across a 500 kb region centered on rs11099601 (chr4: 84,132,874-84,631,193 from GRCh37/hg19) using the March 2012 release of the 1000 Genomes Project as a reference panel. Subsequent association analysis for overall breast cancer risk revealed that rs11099601 was located in a region of approximately 135 kb exhibiting strong LD (Fig. 1). SNP rs11099601 remained one of the most strongly associated SNPs, along with three other perfectly correlated imputed SNPs (r^2 =1.0), namely rs4235062 (P=2.40x10^{-6}), rs6838225 (P=3.70x10^{-6}) and rs13142756 (P=4x10^{-6}) (Fig.1) (S3 Table). 88 SNPs were strongly correlated with rs11099601 (r^2>0.8; S4 Table) and hence not distinguishable as potential causal variants on the basis of association data alone.

Functional annotation of locus 4q21

In order to identify potential candidate causal variants at the 4q21 locus, we overlaid the associated variants with publicly available functional annotations. The analysis was performed on the subset of 88 variants strongly correlated with the lead SNP, rs11099601 (r^2>0.8). We first performed analyses using RegulomeDB (http://www.regulomedb.org) in order to obtain a predicted score of functionality for the set of variants. Interestingly, variant rs11099601 was one of three variants with the highest scores, along with rs1494961 and rs6535481. The corresponding RegulomeDB score (1f) (S4 Table) suggests that these variants are likely to affect transcription factor binding and to be linked to expression of a target gene. The scores for the other three strongest associated SNPs, namely rs4235062, rs6838225 and rs13142756, were not suggestive of functionality (S4 Table – for a description of the RegulomeDB scoring scheme and

referenced datatypes refer to http://www.regulomedb.org). Five other highly correlated SNPs (rs10008742, rs6844460, rs7691492, rs526064, rs813298), however, also had high scores (2b), albeit lower than that of the lead SNP rs11099601, indicative of likely affecting transcription factor binding.

We then analysed ENCODE chromatin biofeatures, namely DNase I hypersensitivity, chromatin state segmentation by HMM (chromHMM) and histone modifications of epigenetic markers H3K4, H3K9 and H3K27 in all breast cell lines available in ENCODE, including breast myoepithelial cells, HMEC mammary cell line, and breast cancer cell line MCF-7. Analysis of these biofeatures revealed an overlap between H3K9Ac, a histone mark associated with active promoters, and our candidate variant, rs11099601 in breast myoepithelial cells. Further analysis of other genotyped and imputed variants correlated with rs11099601, revealed that only rs6844460 (P=4.2x10^{-6}, r^2=0.967) overlapped with several chromatin biofeatures in mammary cells. rs6844460, which is located within intron 1 of *FAM175A,* overlapped with a DNase hypersensitivity site in MCF-7 cells, with H3K4me3 histone marks (associated with active promoters) in breast myoepithelial cells, HMEC and MCF-7 cell lines, with H3K9Ac histone marks in both breast myoepithelial cells and HMEC cells, and with H3K27Ac histone marks in HMEC. ChromHMM data also predicts that this variant lies within an active promoter region in breast cell lines (Fig. 2A). Moreover, rs6844460 overlapped with a binding site for transcription factor Max (MYC Associated Factor X) in MCF7 cells.

In order to identify potential target genes, we analysed enhancer-promoter interactions using ChiA-PET data for CCCTC-binding factor (CTCF) and DNA polymerase II (PolII) in MCF-7 breast tumour derived cells. Multiple, dense, chromosomal interactions were observed in ChiA-PET data for PolII across most of the entire locus, especially in the region encompassing rs11099601, in the vicinity of the promoter regions of *HELQ, MRPS18C* and *FAM175A* genes.

ChiA-PET data for CTCF in MCF-7 cells showed fewer interactions, none of which encompassed variant rs11099601. Similarly Hi-C data revealed few interactions in HMEC cells, none of which included our top candidate SNP (Fig. 2B).

Lastly, although super-enhancers mapped to the 4q21 locus in HMEC mammary cells, none overlapped with our top candidate SNPs (Fig. 2C). Predicted enhancer-promoter interactions were observed with the promoters of *AGPAT9*, *COQ2*, *HELQ* and *MRPS18C* genes in HMEC cells. However amongst these, only interactions with *MRPS18C* overlapped with our top putative candidate functional variants (rs11099601 and rs6844460) (Fig. 2C).

Analysis of RNASeq data from ENCODE showed high levels of expression for *MRPS18C* in both HMEC and MCF-7 while *HELQ* and *FAM175A* are expressed at very low levels in these cell lines (Fig. 2D). However, as illustrated in Fig. 3, analysis of TCGA breast cancer RNAseq data in primary tumor (n=765), adjacent normal (n=93) and metastasis (n=6) showed that *HELQ*, *FAM175A* and *HPSE*, but not *MRPS18C*, were all found to be differentially expressed between normal breast and tumor tissue (P=1x10^{-45}, P=6.6x10^{-31}, P=7.3x10^{-10}, and P=0.28, respectively, as determined by a Kruskal-Wallis rank sum test). Further analysis comparing the tumor expression levels of these genes between the 5 molecular subtypes of breast cancer, namely: Luminal A, Luminal B, Her2-enriched, Basal-like and Normal-like, showed that while *HELQ* and *FAM175A* expression levels are decreased in Basal-like tumors (P=1.3x10^{-18} and P=3.5x10^{-36}, respectively (Kruskal-Wallis test), *MRPS18C* and *HPSE* expression were found to be up regulated in Basal-like carcinomas (P=1.2x10^{-5}, P=1.6x10^{-33}) (Fig. 4).

Expression Quantitative Trait Locus Analysis (eQTL) in breast tissue

In order to identify associations between candidate variants and expression levels of genes within the 4q21 region, we analyzed all genotyped and imputed SNPs within a 1Mb region centered

around the most significant SNP (rs11099601), in normal and breast cancer tissue. Significant eQTL associations were observed for numerous SNPs in the region in both normal breast and tumors (Fig. 5). In the breast cancer tissue dataset BC241, the most strongly expression-associated SNP at this locus was our top risk SNP rs11099601, which was associated with expression levels of *HELQ*, (with $P=8.28\text{x}10^{-14}$ and $r^2=0.20$, where the r^2 value indicates the percentage of variance in *HELQ* expression levels explained by rs11099601) (Fig. 6A). A decrease in *HELQ* expression levels was observed with increasing copy number of the rs11099601 (C) allele (Fig. 6A). Multiple SNPs within the 1 Mb region were also associated with expression of *HELQ*, all of which were correlated with rs11099601 ($r^2>0.3$). No significant eQTLs were observed between rs11099601 and other genes in this region, namely *COQ2*, *HPSE*, *MRPS18C*, *FAM175A*, or *AGPAT9*, using data from the BC241 sample set.

In the TCGA BC765 breast cancer dataset, *HELQ* expression levels were not associated with rs11099601 ($P=0.34$ and $r^2=0.00099$) (Fig. 6B) or with any other SNPs in this region. Weak associations were only observed between rs11099601 and expression levels for *MRPS18C* ($P=1.25\text{x}10^{-4}$ and $r^2=0.02$) (Fig. 6F) and *FAM175A* ($P=3.83\text{x}10^{-3}$ and $r^2=0.011$) (Fig.6H).

As *HELQ* and *MRPS18C* can be transcribed into several different isoforms, further isoform-specific analysis was performed in the TCGA BC765 breast cancer dataset. Indeed, in contrast to the expression data generated from the Norwegian sample sets, which were obtained using expression arrays, expression data from the TCGA datasets used in the current study were obtained by RNA-Sequencing, thus allowing further analysis of different gene isoforms. Thus, in the BC765 dataset, these analyses resulted in the identification of significant eQTLs for an isoform of *HELQ* (uc101ikb) ($P=2.71\text{x}10^{-11}$ and $r^2=0.056$) (Fig. 6C), corresponding to a long isoform of the gene with one exon lacking. These analyses also further revealed highly significant associations for the *MRPS18C* isoform uc003hor ($P=1.94\text{x}10^{-27}$ and $r^2=0.143$) (Fig. 6G).

Similar to what is observed in the TCGA BC765 breast cancer dataset, gene-normalized analysis in the TCGA normal breast tissue dataset NB93 did not reveal significant associations between *HELQ* expression levels and rs11099601 (P=0.15 and r^2=0.017) (Fig. 6D) while isoform-normalized analysis showed associations with *HELQ* isoform uc101ikb (P=9.90x10^{-05} and r^2=0.153) (Fig. 6E).

In normal breast tissue from the NBCS (NB116), the strongest eQTLs were observed for *HPSE*, where rs11099601 was associated with a decrease in *HPSE* expression levels (P=4.57x10^{-3}, r^2=0.0645) (Fig. 6I). rs11099601 was not associated with the expression levels of any other genes in this region.

Although associations were detected between several genes and our top risk SNP in the different sample sets, a lack of consistency in eQTL associations between the two breast cancer sample sets was observed. It should be noted that expression data were obtained trough different approaches as previously mentioned, i.e expression array (44K Agilent array) for BC241 and RNA-Sequencing for BC765 (Illumina RNAseq). Moreover, there are differences in the overall PAM50 subtype distributions between these two sample sets. As depicted in S3 Fig., differences are noted mainly in the distribution of Luminal A (28.22% in BC241 compared to 49.33% for BC765), Her2 (15.35% in BC241 compared to 8.16% for BC765) and Normal-like (14.52% in BC241 compared to 2.41% for BC765) subtypes. Expression levels of *HELQ*, and of other candidate genes, were shown to vary significantly between these molecular subtypes (Fig. 4) and thus a different distribution of these subtypes between the two sample sets could explain the underlying lack of replication in the eQTL analyses.

Discussion

It is well recognized that genetic variants located in genomic regions that regulate gene expression are major causes of human diversity and may also be important susceptibility factors for complex diseases and traits. Indeed, it has been shown that approximately 30% of expressed genes show evidence of *cis*-regulation by common polymorphic alleles [47]. Moreover, in recent years, GWAS have identified thousands of variants associated with various diseases/traits, ~90% of which localize outside of known protein-coding regions [42, 43], implicating a regulatory role for these variants.

In the present study, we have assessed the association with breast cancer risk of 313 regulatory SNPs in genes involved in the etiology of cancer (see S1 Table for complete list of SNPs and genes), in 46,451 breast cancer cases and 42,599 controls of European ancestry. Using this approach, we identified rs11099601 (OR=1.05, P=5.6x10^{-6}), a novel breast cancer susceptibility locus on chromosome 4q21. Analysis of imputed SNPs across a 500Kb region surrounding rs11099601 revealed that this variant remained one of the strongest risk signals, tagging a set of 76 strongly correlated SNPs across a 135Kb LD block containing several genes, including *COQ2, HPSE, HELQ, MRPS18C, FAM175A (ABRAXAS)* and *AGPAT9.*

Functional annotation of the 4q21 locus with ENCODE biofeatures in mammary cell lines pointed toward rs11099601 as one of the most likely functional variants in this region. eQTL analysis showed significant eQTLs in normal and breast cancer tissue for several variants in the 4q21 region, including rs11099601. The strongest associations for rs11099601 and expression were observed in breast carcinomas for *MRPS18C* and *HELQ* and explain approximately 14%

and 20% of their expression variance, respectively (Fig. 6). Other genes whose expression correlated with this eQTL included *HPSE* and *FAM175A*.

These genes represent interesting candidates for further analyses related to breast cancer susceptibility. Indeed, analysis of TCGA breast cancer RNAseq data showed that *HELQ*, *FAM175A* and *HPSE* were found to be differentially expressed between normal breast and tumor tissue and further analysis showed that *HELQ* and *FAM175A* expression levels are significantly decreased in basal-like tumors.

HELQ is a single-stranded DNA-dependent ATPase and DNA helicase involved in DNA repair and signaling in response to ICL. Genetic disruption of *HELQ* in human cells enhances cellular sensitivity and chromosome radial formation by the ICL-inducing agent mitomycin C (MMC). After treatment with MMC, reduced phosphorylation of CHK1 occurs in knockout cells and accumulation of G2/M cells is reduced [51]. Furthermore, it was recently shown that Helq helicase-deficient mice exhibit subfertility, germ cell attrition, ICL sensitivity, and tumor predisposition [52]. A meta-analysis of 22 GWAS, as well as a recent GWAS involving ~70,000 women performed in the BCAC, have both identified rs4693089, located in an intron of *HELQ* and perfectly correlated with rs11099601, as associated with age at natural menopause ($p=2.4x10^{-19}$ and $p=9.2x10^{-23}$, respectively) [53, 54]. Moreover, a GWAS of upper aero-digestive tract cancers conducted by the International Head and Neck Cancer Epidemiology Consortium identified rs1494961, a missense mutation V306I in the second exon of *HELQ* gene perfectly correlated with rs11099601 ($r^2=1$), to be associated with increased risk of upper aero-digestive tract cancers in their combined analysis ($P=1x10^{-8}$) [55]. Another study by the same group analyzed the role of DNA repair pathways in upper aero-digestive tract cancers [56]. This study showed that the polymerase pathway, to which the *HELQ* gene belongs, is the only pathway significant for all upper aero-digestive tract cancer sites combined and that this association is

entirely explained by the association with rs1494961 (P=2.65×10^{-4}) [56]. Finally, a recent study reported the mutation screening of *HELQ* in 185 Finnish breast or ovarian cancer families [57]. This study did not provide evidence for a role of *HELQ* in breast cancer susceptibility in the Finnish population, but analyses in other populations and larger datasets are needed to further assess its role in breast cancer predisposition [57], especially with regard to the involvement of rare variants. In the current study, we have shown *HELQ* to be differentially expressed between normal breast and tumor tissue and to be significantly down regulated in basal-like breast tumors compared to ER positive tumors, suggesting that altered gene expression levels, potentially mediated through the effect of regulatory variants, could be one of the mechanisms contributing to breast cancer susceptibility. Previous studies have provided some evidence, in known breast cancer susceptibility genes *BRCA1* [58] and *BRCA2* [59], of genetic variants associated with allelic expression differences which could affect the risk of breast cancer in mutation carriers through altering expression levels of the wild-type allele. Also, a recent study showed suggestive associations between DAE associated variants located in breast cancer susceptibility chromosomal regions, and prognosis (*ZNF331* and *CHRAC1*) [60].

Another gene in this locus, *FAM175A*, is involved in DNA damage response and double-strand break (DSB) repair. It is a component of the BRCA1-A complex, acting as a central scaffold protein that assembles the various components of the complex and mediates the recruitment of BRCA1 [61-63]. Further evidence rendering *FAM175A/ABRAXAS* an interesting candidate gene is a recent report showing that both homozygous and heterozygous *Abraxas* knockout mice exhibited decreased survival and increased tumor incidence [64]. This study also showed that somatic deletion of the *ABRAXAS* locus on chromosome 4q21 is found in human ovarian and breast cancers (especially basal subtype), and this loss is well correlated with reduced *ABRAXAS* expression in these cancers [64]. Moreover, Solyom et al. reported a novel germline *ABRAXAS*

mutation (p.Arg361Gln) in Northern Finnish breast cancer families which affects the nuclear localization of the protein and consequently reduces the formation of BRCA1 and Rap80 foci at DNA damage sites, leading to ionizing radiation hypersensitivity of cells and partially impairing the G2/M checkpoint [65]. In the current study, *FAM175A* expression was found to be significantly down-regulated in ER negative breast tumors. Our group has also, in parallel to the present study, conducted a population-based case-control mutation screening study of the coding exons and exon/intron boundaries of *ABRAXAS* in 1250 breast cancer cases and 1250 controls from the Breast Cancer Family Registry, including individuals from different ethnic groups such as Caucasian, Latino, East Asian and African-American ancestry. Although this study did not reveal evidence of association of the identified variants with breast cancer risk, two variants were identified and were shown to diminish the phosphorylation of γ-H2AX, an important biomarker of DNA double-strand breaks [66].

Lastly, *MRPS18C* encodes a protein that belongs to the ribosomal protein S18P family, which includes three proteins (MRPS18A, MRPS18B, MRPS18C) having significant sequence similarity to bacterial S18 proteins. MRPS18C is part of the small subunit (28S) of the mitochondrial ribosome involved in oxidative phosphorylation and thus the role of this protein in breast cancer susceptibility is unclear. It was reported that MRPS18B (MRPS18-2) binds to RB [67] and prevents the formation of the E2F1-RB complex that leads to elevated levels of free E2F1 protein in the nucleus and the subsequent promotion of S phase entry [68]. Overexpression of human MRPS18B caused transformation of terminally differentiated rat skin fibroblasts and transformed cells became tumorigenic in SCID (severe combined immunodeficiency) mice [69]. These transformed cells showed anchorage-independent growth and loss of contact inhibition; they expressed epithelial markers, showed increased telomerase activity, disturbance of the cell

cycle, and chromosomal instability, leading the authors to suggest that MRPS18B is a newly identified oncoprotein [69]. Although these results suggest that MRPS18B may be involved in carcinogenesis, there is currently no evidence showing that MRPS18C is involved in processes other than oxidative phosphorylation.

Conclusion

Phenotypic differences among cell types, individuals, and populations are determined by variation in gene expression, a substantial proportion of which is driven by genetic variants residing in regulatory elements near the affected genes. Analysis of variants associated with differential allelic expression has allowed us to identify a novel locus on chromosome 4q21 associated with breast cancer risk. Subsequent tissue specific eQTL analyses have confirmed significant eQTLs for this locus in both normal and breast cancer tissue.

At the time of study design, data on differential allelic expression was not available in breast tissue, leading us to perform the selection of candidate variants in other cell types such as lymphoblastoid cell lines, fibroblasts and monocytes. This constitutes a limitation of our study which may explain why some of the associations observed between the selected variants and DAE in these cells types were not replicated in the eQTL analyses performed in normal breast and/or breast cancer cells. Indeed, SNPs associated with variation in gene expression have now been mapped for a variety of tissues, highlighting their tissue dependent properties and the need for expression profiling of a diverse panel of cell types.

Hence, further functional characterization of the 4q21 locus, and replication in a larger dataset, would be relevant to provide more robust evidence of the involvement of this region in breast cancer susceptibility as well as identify the gene(s) and biological mechanism(s) underlying this susceptibility.

Materials and Methods

Sample Selection

A total of 46,451 breast cancer cases and 42,599 controls of European ancestry were included from 41 studies participating in the Breast Cancer Association Consortium (BCAC). Studies were population-based or hospital-based case-control studies, including nested case-control studies within cohorts. Some studies selected cases by age, or oversampled cases with a family history (S5 Table). Studies provided ~2% of samples in duplicate for quality control purposes (see below). Study subjects were recruited on protocols approved by the Institutional Review Boards at each participating institution, and all subjects provided written informed consent.

SNP Selection

SNP selection was performed by first identifying a list of genes of interest, which was determined by the involvement of these genes in cancer related pathways and/or mechanisms. The list of genes was established by researching published results and/or by using available public databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/). Thereafter, DEA SNPs falling within these gene regions were identified using previously reported data on allelic expression *cis*-associations, derived using: 1) the llumina Human1M-duo BeadChip for lymphoblastoid cell lines from Caucasians (CEU population) (n=53) [47], the Illumina Human 1M Omni-quad for primary skin fibroblasts derived from Caucasian donors (n = 62) [49, 70], and the Illumina Infinium II assay with Human 1.2 M Duo custom BeadChips v1 for human primary monocytes (n=188) [71]. Briefly, 1000 Genomes project data was used as a reference set (release 1000G Phase I v3) for the imputation of genotypes from HapMap individuals. Untyped markers were inferred using algorithms implemented in IMPUTE2. The

unrelated fibroblast panel consisted of 31 parent-offspring trios, where the genotypes of offspring were used to allow for accurate phasing. Mapping of each allelic expression trait was carried out by first normalizing allelic expression ratios at each SNP using a polynomial method [72] and then calculating averaged phased allelic expression scores across annotated transcripts, followed by correlation of these scores to local (transcript +/-500 kb) SNP genotypes in fibroblasts as described earlier [70].

Three hundred fifty-five genetic variants were selected on the basis of evidence of association with DAE in 175 genes involved in cancer-related pathways as described above (see S1 Table for complete list of SNPs and genes). Following selection, SNPs were submitted for design and inclusion on a custom Illumina Infinium array (iCOGS), as part of a BCAC genotyping initiative (see Genotyping and Quality Control section below). After undergoing design and post-genotyping quality control, 313 SNPs remained for analysis.

Genotyping and Quality Control

Genotyping was carried out as part of a collaboration between BCAC and three other consortia (the Collaborative Oncological Gene-environment Study, COGS). Full details of SNP selection, array design, genotyping and post-genotyping quality control (QC) have been published [35]. Briefly, three categories of SNPs were chosen for inclusion on the array: (i) SNPs selected on the basis of pooled GWAS data, (ii) SNPs selected for the fine-mapping of published risk loci and (iii) candidate SNPs selected on the basis of previous analyses or specific hypotheses. The 313 SNPs described in the current study were candidate SNPs selected on the basis of the hypothesis that regulatory variants are involved in breast cancer susceptibility. In general, only SNPs with an Illumina design score of 0.8 or greater were considered. SNPs were preferentially accepted if they had a design score of 1.1 (i.e. had previously been genotyped on an Illumina platform). If

not, we sought SNPs with r^2=1 with the selected SNP, and selected the SNP with the best design score. If no such SNP was available, we selected SNPs with r^2>0.8 with the chosen SNP, and selected the SNP with the best design score. For the COGS project overall, genotyping of 211,155 SNPs in samples was conducted using a custom Illumina Infinium array (iCOGS) in four centers. Genotypes were called using Illumina's proprietary GenCall algorithm. Standard quality control measures were applied across all SNPs and all samples genotyped as part of the COGS project [35]. After quality control, genotype data were available for 48 155 breast cancer cases and 43 612 controls, and call rates for all SNPs were >95%.

Statistical Analysis

Per-allele log-odds ratios (ORs) were estimated using logistic regression, adjusted for principal components and study, as described previously [35]. *P*-values were estimated using Wald test. For imputation, genotype data from 48,155 breast cancer cases and 43,612 controls were used to estimate genotypes for other common variants across a 500 kb region on chromosome 4 (chr4: 84,132,763-84,632,763 - NCBI build 37 assembly), with IMPUTE v.2.2 and the March 2012 release of the 1,000 Genomes Project as reference panel. In all analyses, only SNPs with imputation information/accuracy r2>0.30 were considered [40].

Linkage Disequilibrium

LD values were computed using 118 independent individuals from the CEU population of the 1,000 Genome dataset (v3, release 20110521, downloaded from 1000genomes.ebi.ac.uk on April 2013) [73]. The relevant subset was extracted from the raw data using VCFtools (v0.1.7) [74] and the paired r^2 statistics were obtained for all target loci using PLINK! (v1.07) [75]. The linkage heatmaps and the association plots were produced on the R platform (v3.0) using the package LDheatmap [76].

Breast cancer association analyses performed in *BRCA1* and *BRCA2* mutation carriers.

Associations with breast cancer risk were evaluated within a retrospective cohort framework, by modelling the retrospective likelihood of the observed genotypes conditional on the disease phenotype. These analyses are described in detail elsewhere [77, 78].

Functional annotation

Two publicly available tools, RegulomeDB [79] and HaploReg V4 [80], were also used to evaluate candidate variants. For a full description of the RegulomeDB scoring scheme refer to (http://www.regulomedb.org).

Publicly available genomic data was also used to annotate each SNP most strongly associated with breast cancer risk at locus 4q21 (for data sources refer to S6 Table). The following regulatory features were obtained for breast cell types from ENCODE and NIH Roadmap Epigenomics data through the UCSC Genome Browser: DNase I hypersensitivity sites, Chromatin Hidden Markov Modelling (ChromHMM) states, histone modifications of epigenetic markers more specifically commonly used marks associated with enhancers (H3K4Me1 and H3K27Ac) and promoters (H3K4Me3 and H3K9Ac), and transcription factor ChiP-seq data.

To identify putative target genes, we examined potential functional chromatin interactions between distal and proximal regulatory transcription-factor binding sites and the promoters at the risk loci, using the Chromatin Interaction Analysis by Paired End Tag (ChiA-PET) and Genome conformation capture (Hi-C, 3C and 5C) datasets downloaded from GEO.

Maps of active mammary super-enhancer regions in HMEC cells were obtained from Hnisz et al. [81]. Predicted enhancer-promoter determined interactions were obtained from the integrated method for predicting enhancer targets (IM-PET) described in He et al. [82].

RNA-Seq data from ENCODE was used to evaluate the expression of exons across the 4q21 locus in HMEC and MCF7 cell lines. For HMEC and MCF7, alignment files from 4 and 19 expression datasets respectively were downloaded from ENCODE using a rest API wrapper (ENCODExplorer R package) [83] in the bam format and processed using metagene R packages [84] to normalize in Reads per Millions aligned, and to convert in coverages.

eQTL analyses

The influence of germline genetic variations on gene expression was assessed using a linear regression model, as implemented in the R library eMAP (http://www.bios.unc.edu/~weisun/software.htm). An additive effect was assumed by modeling subjects' copy number of the rare allele, i.e. 0, 1 or 2 for a given genotype. Only relationships in *cis* (defined as those in which the SNP resided less than 1 MB up or down from the center of the transcript) were investigated. eQTL analyses were performed on both normal breast and tumor tissues, and included the following materials: Normal Breast: NB116 (n=116) consists of samples from women of Caucasian ancestry recruited in Oslo, comprising expression data from normal breast biopsies (n=73), reduction plastic surgery (n=34) and adjacent normal (n=9) (adjacent to tumour). Genotyping was performed using the iCOGS SNP array, and gene expression levels were measured with Agilent 44K [85, 86]. NB93 is the Caucasian fraction of the TCGA dataset for which adjacent normal breast expression data were available, n=93 for the data normalized per gene, and n=94 for the data normalized per isoform. Birdseed processed germline genotype data from the Affy6 SNP array were obtained from the TCGA dbGAP data portal [87]. Gene expression levels were assayed by RNA sequencing, RSEM (RNAseq by Expectation-Maximization, [88] normalized both per gene and per isoform, as obtained from the TCGA consortium [87]. The data was log2 transformed, and unexpressed genes were excluded prior to

eQTL analysis. Breast carcinomas: BC241, is a Caucasian sample set recruited from Oslo, n=241. The sample set includes all stages of breast cancer, and genotypes were obtained with the iCOGS SNP array, and mRNA expression data was from the Agilent 44K array [86, 89]. BC765 comprises samples from the TCGA breast cancer sample set of Caucasian origin [87], n=765 for the data normalized per gene, and n=766 for the data normalized per isoform. Genotyping platform was Affy6, and gene expression was measured using RNA sequencing. See NB93 for a more detailed description. For all sample sets, the genotyping data was processed as follows: SNPs with call rates <0.95 or minor allele frequencies < 0.05 were excluded, as were SNPs out of Hardy Weinberg equilibrium with $P<10^{-13}$. All samples with a call rate below 80% were excluded. Identity by state was computed using the R GenABEL package [90], and closely related samples with IBS>0.95 were removed. The SNP and sample filtration criteria were applied iteratively until all samples and SNPs met the stated thresholds.

Acknowledgments

BCAC thanks all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. This study would not have been possible without the contributions of the following: Per Hall (COGS); Douglas F. Easton, Paul Pharoah, Kyriaki Michailidou, Manjeet K. Bolla, Qin Wang (BCAC), Andrew Berchuck (OCAC), Rosalind A. Eeles, Douglas F. Easton, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Georgia Chenevix-Trench, Antonis Antoniou, Lesley McGuffog, Fergus Couch and Ken Offit (CIMBA), Joe Dennis, Alison M. Dunning, Andrew Lee, and Ed Dicks, Craig Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory, Javier Benitez, Anna Gonzalez-Neira and the staff of the CNIO genotyping unit, Jacques Simard and Daniel C. Tessier, Francois Bacot, Daniel Vincent, Sylvie LaBoissière and Frederic Robidoux and the staff of the McGill University and Génome Québec Innovation Centre, Stig E. Bojesen, Sune F. Nielsen, Borge G. Nordestgaard, and the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility. ABCFS thanks Maggie Angelakos, Judi Maskiell, and Gillian Dite. ABCS thanks Sten Cornelissen, Richard van Hien, Linde Braaf, Frans Hogervorst, Senno Verhoef, Laura van't Veer, Emiel Rutgers, C Ellen van der Schoot, Femke Atsma. BBCS Acknowledges Eileen Williams, Elaine Ryder-Mills, and Kara Sargus. BIGGS Acknowledges Niall McInerney, Gabrielle Colleran, Andrew Rowan, and Angela Jones. BSUCH thanks Peter Bugert, Medical Faculty Mannheim. CGPS thanks Staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. The

Danish Cancer Biobank is acknowledged for providing infrastructure for the collection of blood samples for the cases. CNIO-BCS thanks Guillermo Pita, Charo Alonso, Daniel Herrero, Nuria Álvarez, Pilar Zamora, Primitiva Menendez, and the Human Genotyping-CEGEN Unit (CNIO). CTS acknowledges The CTS Steering Committee includes Leslie Bernstein, Susan Neuhausen, James Lacey, Sophia Wang, Huiyan Ma, Yani Lu, and Jessica Clague DeHart at the Beckman Research Institute of City of Hope, Dennis Deapen, Rich Pinder, Eunjung Lee, and Fred Schumacher at the University of Southern California, Pam Horn-Ross, Peggy Reynolds, Christina Clarke Dur and David Nelson at the Cancer Prevention Institute of California, and Hoda Anton-Culver, Argyrios Ziogas, and Hannah Park at the University of California Irvine. ESTHER thanks Hartwig Ziegler, Sonja Wolf, Volker Hermann, Christa Stegmaier, Katja Butterbach. GENICA Thanks The GENICA Network: Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany [HB, Wing-Yee Lo, Christina Justenhoven], German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ) [HB], Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany [Yon-Dschun Ko, Christian Baisch], Institute of Pathology, University of Bonn, Germany [Hans-Peter Fischer], Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany [Ute Hamann], Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany [TB, Beate Pesch, Sylvia Rabstein, Anne Lotz]; and Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany [Volker Harth]. HEBCS thanks Sofia Khan, Taru A. Muranen, Kristiina Aittomäki, Kirsimari Aaltonen, Karl von Smitten, Irja Erkkilä. The HMBCS study thanks Peter Hillemanns, Hans Christiansen and Johann H. Karstens. KBCP thanks Eija Myöhänen, Helena Kemiläinen. kConFab/AOCS wish to thank Heather Thorne,

Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab. LMBC acknowledges Gilian Peuteman, Dominiek Smeets, Thomas Van Brussel and Kathleen Corthouts. MBCSG (Milan Breast Cancer Study Group) thanks Paolo Peterlongo of IFOM, the FIRC Institute of Molecular Oncology; Siranoush Manoukian, Bernard Peissel, Daniela Zaffaroni and Jacopo Azzollini of the Fondazione IRCCS Istituto Nazionale dei Tumori (INT); Monica Barile and Irene Feroce of the Istituto Europeo di Oncologia (IEO), and the personnel of the Cogentech Cancer Genetic Test Laboratory. MTLGEBCS would like to thank Martine Tranchant (CHU de Québec Research Center), Marie-France Valois, Annie Turgeon and Lea Heguy (McGill University Health Center, Royal Victoria Hospital; McGill University) for DNA extraction, sample management and skillful technical assistance. J.S. is Chairholder of the Canada Research Chair in Oncogenetics. The following are NBCS Collaborators that we would like to thank: Dr. Kristine K.Sahlberg, PhD (Department of Research, Vestre Viken Hospital, Drammen, Norway and Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway), Dr. Lars Ottestad, MD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway), Prof. Em. Rolf Kåresen, MD (Institute of Clinical Medicine, University of Oslo, Oslo, Norway and Department of Breast- and Endocrine Surgery, Division of Surgery, Cancer and Transplantation, Oslo University Hospital, Oslo, Norway), Dr. Anita Langerød, PhD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway), Dr. Ellen Schlichting, MD (Section for Breast- and Endocrine Surgery, Department of Cancer, Division of Surgery, Cancer and Transplantation Medicine, Oslo

University Hospital, Oslo, Norway), Dr. Marit Muri Holmen, MD (Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway), Prof. Toril Sauer, MD (Department of Pathology at Akershus University hospital, Lørenskog, Norway and Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway), Dr. Vilde Haakensen, MD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway), Dr. Olav Engebråten, MD (Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, Department of Oncology, Division of Surgery and Cancer and Transplantation Medicine, Oslo University Hospital, Oslo, Norway and Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway), Prof. Bjørn Naume, MD (Department of Oncology, Division of Surgery and Cancer and Transplantation Medicine, Oslo University Hospital-Radiumhospitalet, Oslo, Norway and K.G. Jebsen Centre for Breast Cancer, Institute for Clinical Medicine, University of Oslo, Oslo, Norway.), Dr. Cecile E. Kiserud, MD (National Advisory Unit on Late Effects after Cancer Treatment, Department of Oncology, Oslo University Hospital, Oslo, Norway and Department of Oncology, Oslo University Hospital, Oslo, Norway), Dr. Kristin V. Reinertsen, MD (National Advisory Unit on Late Effects after Cancer Treatment, Department of Oncology, Oslo University Hospital, Oslo, Norway and Department of Oncology, Oslo University Hospital, Oslo, Norway), Assoc. Prof. Åslaug Helland, MD (Department of Genetics, Institute for Cancer Research and Department of Oncology, Oslo University Hospital Radiumhospitalet, Oslo, Norway), Dr. Margit Riis, MD (Dept of Breast- and Endocrine Surgery, Oslo University Hospital, Ullevål, Oslo, Norway), Dr. Ida Bukholm, MD (Department of Breast-Endocrine Surgery, Akershus University Hospital, Oslo, Norway and Department of Oncology, Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital, Oslo, Norway), Prof. Per Eystein Lønning, MD (Section of Oncology, Institute of Medicine, University of Bergen and Department of Oncology,

Haukeland University Hospital, Bergen, Norway), OSBREAC (Oslo Breast Cancer Research Consortium), Prof. Anne-Lise Børresen-Dale, PhD (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway and Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Norway) and Grethe I. Grenaker Alnæs, M.Sc. (Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway). NBHS thanks study partcipants and research staff for their contributions and commitment to this study. OBCS thanks Arja Jukkola-Vuorinen, Mervi Grip, Saila Kauppila, Meeri Otsukka and Kari Mononen for their contributions to this study. OFBCR thanks Teresa Selander, Nayana Weerasooriya. ORIGO thanks E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The LUMC survival data were retrieved from the Leiden hospital-based cancer registry system (ONCDOC) with the help of Dr. J. Molenaar. PBCS thanks Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao and Michael Stagner. pKARMA acknowledges the Swedish Medical Research Counsel. RBCS thanks Petra Bos, Jannet Blom, Ellen Crepin, Elisabeth Huijskens, Annette Heemskerk, and the Erasmus MC Family Cancer Clinic. SASBAC thanks The Swedish Medical Research Counsel. SBCS thanks Sue Higham, Helen Cramp, Ian Brock, Sabapathy Balasubramanian, Malcolm W.R. Reed and Dan Connley. SEARCH thanks The SEARCH and EPIC teams. SGBCC would like to thank the participants and research coordinator Kimberley Chua. SKKDKFZS acknowledges all study participants, clinicians, family doctors, researchers and technicians for their contributions and commitment to this study. UKBGS wants to thank Breast Cancer Now and the Institute of Cancer Research for support and funding of the Breakthrough Generations Study, and the study participants, study staff, and the doctors, nurses and other health care providers and health

information sources who have contributed to the study. We acknowledge NHS funding to the Royal Marsden/ICR NIHR Biomedical Research Centre.

The authors would like to acknowledge Dr. Katherine A. Hoadley for normalization and sharing of TCGA BRCA RNAseq gene expression data.

Disclosure of potential conflicts of interest

We do not expect any conflict of interests to disclose. If the manuscript moves to the revision stage, we will promptly coordinate the completion of the Conflict of Interest forms for all co-authors and provide these to the editorial office.

Funding

This work was supported by the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program – grant # CRN-87521, the Ministry of Economy Innovation and Exportations– grant # PSR-SIIRI-701 and by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, and the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Genome Québec and The Quebec Breast Cancer Foundation for the PERSPECTIVE project. Jacques Simard is Chairholder of the Canada Research Chair in Oncogenetics. Silje Nord is financed by a carrier grant from the Norwegian Regional Health authorities (Grant number 2014061). S.N. is a researcher on a carrier grant from the South-Eastern Norway Regional Health Authority (Grant number 2014061).

BCAC is funded by Cancer Research UK [C1287/A10118, C1287/A12014] and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the

National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]; BBMRI-NL, which is a Research Infrastructure financed by the Dutch government (NWO 184.021.007); and the Dutch National Genomics Initiative. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCS is funded by Cancer Research UK and Breast Cancer Now and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). ES is supported by NIHR Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London, United Kingdom. IT is supported by the Oxford Biomedical Research Center. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). The CECILE study was funded by Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Ligue contre le Cancer Grand Ouest, Agence Nationale de Sécurité Sanitaire (ANSES), Agence Nationale de la Recherche (ANR). The CGPS was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council and Herlev Hospital. The CNIO-BCS was supported by the Instituto de Salud Carlos III, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española

Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI12/00070). The CTS was initially supported by the California Breast Cancer Act of 1993 and the California Breast Cancer Research Fund (contract 97-10500) and is currently funded through the National Institutes of Health (R01 CA77398). Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885. HAC receives support from the Lon V Smith Foundation (LVS39420). The ESTHER study was supportd by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe). The GENICA was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, as well as the Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany. The HEBCS was financilly supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (266528), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation. The HMBCS was supported by a grant from the Friends of Hannover Medical School and by the Rudolf Bartling Foundation. Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and and Bert von Kantzows foundation. The KBCP was financially supported by the special Government Funding (EVO) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer

Organizations, and by the strategic funding of the University of Eastern Finland. kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. Financial support for the AOCS was provided by the United States Army Medical Research and Materiel Command [DAMD17-01-1-0729], Cancer Council Victoria, Queensland Cancer Fund, Cancer Council New South Wales, Cancer Council South Australia, The Cancer Foundation of Western Australia, Cancer Council Tasmania and the National Health and Medical Research Council of Australia (NHMRC; 400413, 400281, 199600). G.C.T. and P.W. are supported by the NHMRC. RB was a Cancer Institute NSW Clinical Research Fellow. LMBC is supported by the 'Stichting tegen Kanker' (232-2008 and 196-2010). Diether Lambrechts is supported by the FWO and the KULPFV/10/016-SymBioSysII. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MBCSG (Milan Breast Cancer Study Group) is supported by grants from the Italian Association for Cancer Research (AIRC) and by funds from the Italian citizens who allocated the 5/1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale Tumori, according to Italian laws (INT-Institutional strategic projects “5x1000”). The MCBCS was supported by the NIH grants CA192393, CA116167, CA176785 an NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation and a generous gift from the David F. and Margaret T. Grohne Family Foundation. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by

infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR) and the Australian Institute of Health and Welfare (AIHW), including the National Death Index and the Australian Cancer Database. The MEC was support by NIH grants CA63464, CA54281, CA098758 and CA132839. The work of MTLGEBCS was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701. The NBCS has received funding from the K.G. Jebsen Centre for Breast Cancer Research; the Research Council of Norway grant 193387/V50 (to A-L Børresen-Dale and V.N. Kristensen) and grant 193387/H10 (to A-L Børresen-Dale and V.N. Kristensen), South Eastern Norway Health Authority (grant 39346 to A-L Børresen-Dale) and the Norwegian Cancer Society (to A-L Børresen-Dale and V.N. Kristensen). The OBCS was supported by research grants from the Finnish Cancer Foundation, the Academy of Finland (grant number 250083, 122715 and Center of Excellence grant number 251314), the Finnish Cancer Foundation, the Sigrid Juselius Foundation, the University of Oulu, the University of Oulu Support Foundation and the special Governmental EVO funds for Oulu University Hospital-based research activities. The Ontario Familial Breast Cancer Registry (OFBCR) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human

Services, USA. The pKARMA study was supported by Märit and Hans Rausings Initiative against Breast Cancer. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318). The SASBAC study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. The SBCS was supported by Yorkshire Cancer Research S295, S299, S305PA and Sheffield Experimental Cancer Medicine Centre. SEARCH is funded by a programme grant from Cancer Research UK [C490/A10124] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. SEBCS was supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0000347). SGBCC is funded by the NUS start-up Grant, National University Cancer Institute Singapore (NCIS) Centre Grant and the NMRC Clinician Scientist Award. Additional controls were recruited by the Singapore Consortium of Cohort Studies-Multi-ethnic cohort (SCCS-MEC), which was funded by the Biomedical Research Council, grant number: 05/1/21/19/425. SKKDKFZS is supported by the DKFZ. The SZBCS was supported by Grant PBZ_KBN_122/P05/2004. The TNBCC was supported by: a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation. The UKBGS is funded by Breast Cancer Now and the Institute of Cancer Research (ICR), London. ICR acknowledges NHS funding to the NIHR Biomedical Research Centre.

References

1. Feunteun J, Lenoir GM. BRCA1, a gene involved in inherited predisposition to breast and ovarian cancer. Biochim Biophys Acta. 1996; 1242: 177-180.

2. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, Bell R, Rosenthal J, Hussey C, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994; 266: 66-71.

3. Tavtigian SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, Belanger C, Bell R, Berry S, et al. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. Nat Genet. 1996; 12: 333-337.

4. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995; 378: 789-792.

5. Børresen AL, Andersen T, Garber J, Barbier-Piraux N, Thorlacius S, Eyfjörd J, Ottestad L, Smith-Sørensen B, Hovig E, Malkin D, Friend SH. Screening for germline TP53 mutations in breast cancer patients. Cancer Res. 1992; 52: 3234-3236.

6. Meijers-Heijboer H, Wijnen J, Vasen H, Wasielewski M, Wagner A, Hollestelle A, Elstrodt F, Van den Bos R, De Snoo A, Fat GT, Brekelmans C, Jagmohan S, Franken P, et al. The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype. Am J Hum Genet. 2003; 72: 1308-1314.

7. Le Calvez-Kelm F, Lesueur F, Damiola F, Vallée M, Voegele C, Babikyan D, Durand G, Forey N, McKay-Chopin S, Robinot N, Nguyen-Dumont T, Thomas A, Byrnes GB, et al. Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer

susceptibility: results from a breast cancer family registry case-control mutation-screening study. Breast Cancer Res. 2011; 13: R6.

8. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. 2006; 38: 873-875.

9. Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, et al. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. Am J Hum Genet. 2009; 85: 427-446.

10. Erkko H, Xia B, Nikkilä J, Schleutker J, Syrjäkoski K, Mannermaa A, Kallioniemi A, Pylkäs K, Karppinen SM, Rapakko K, Miron A, Sheng Q, Li G, et al. A recurrent mutation in PALB2 in Finnish cancer families. Nature. 2007; 446: 316-319.

11. Tischkowitz M, Capanu M, Sabbaghian N, Li L, Liang X, Vallée MP, Tavtigian SV, Concannon P, Foulkes WD, Bernstein L, WECARE Study Collaborative Group, Bernstein JL, Begg CB. Rare germline mutations in PALB2 and breast cancer risk: A population-based study. Hum Mutat. 2012; 33: 674-680.

12. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet. 2007; 39: 165-167.

13. Antoniou AC, Foulkes WD, Tischkowitz M. Breast-cancer risk in families with mutations in PALB2. N Engl J Med. 2014; 371: 1651-1652.

14. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans DG, Chenevix-Trench G, et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 2015; 372: 2243-2257.

15. Damiola F, Pertesi M, Oliver J, Le Calvez-Kelm F, Voegele C, Young EL, Robinot N, Forey N, Durand G, Vallée MP, Tao K, Roane TC, Williams GJ, et al. Rare key functional domain missense substitutions in MRE11A, RAD50, and NBN contribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study. Breast Cancer Res. 2014; 16: R58.

16. Heikkinen K, Karppinen SM, Soini Y, Mäkinen M, Winqvist R. Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility. J Med Genet. 2003; 40: e131.

17. Heikkinen K, Rapakko K, Karppinen SM, Erkko H, Knuutila S, Lundán T, Mannermaa A, Børresen-Dale AL, Borg A, Barkardottir RB, Petrini J, Winqvist R. RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability. Carcinogenesis. 2006; 27: 1593-1599.

18. Young E, Feng B-J, Stark A, Damiola F, Durand G, Forey N, Francy TC, Gammon A, Kohlmann WK, Kaphingst KA, McKay-Chopin S, Nguyen-Dumont T, Oliver J, et al. Multigene Testing of Moderate Risk Genes: Be Mindful of the Missense. J Med Genet. 2016; First published on January 19, 2016 doi: 101136/jmedgenet-2015-103398 [Epub ahead of print].

19. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S, Lissowska J, Brinton L, Peplonska B, et al. A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet. 2007; 39: 352-358.

20. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007; 447: 1087-1093.

21. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007; 39: 870–874.

22. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer. Nat Genet. 2007; 39: 865–869.

23. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor–positive breast cancer. Nat Genet. 2008; 40: 703–706.

24. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, Eccles D, Evans DG, Fletcher O, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet. 2009; 41: 585–590.

25. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, Gu K, Fair AM, Cai Q, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet. 2009; 41: 324–328.

26. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet. 2009; 41: 579–584.

27. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, Hughes D, Warren-Perry M, Tapper W, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet. 2010; 42: 504–507.

28. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T, Ghoussaini M, Barrowdale D, EMBRACE, 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. 2010; 42: 885–892.

29. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, Zelenika D, Gut I, Heath S, Palles C, Coupland B, Broderick P, Schoemaker M, et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst. 2011; 103: 425–435.

30. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB, Baglietto L, Balleine R, Bandera EV, et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer. Nat Genet. 2011; 43: 1210–1214.

31. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, Dennis J, Wang Q, Humphreys MK, Luccarini C, Baynes C, Conroy D, Maranian M, et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet. 2012; 44: 312–318.

32. Siddiq A, Couch FJ, Chen GK, Lindström S, Eccles D, Millikan RC, Michailidou K, Stram DO, Beckmann L, Rhie SK, Ambrosone CB, Aittomäki K, Amiano P, et al. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. Hum Mol Genet. 2012; 21: 5373–5384.

33. French JD, Ghoussaini M, Edwards SL, Meyer KB, Michailidou K, Ahmed S, Khan S, Maranian MJ, O'Reilly M, Hillman KM, Betts JA, Carroll T, Bailey PJ, et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am J Hum Genet. 2013; 92: 489-503.

34. Meyer KB, O'Reilly M, Michailidou K, Carlebur S, Edwards SL, French JD, Prathalingham R, Dennis J, Bollan MK, Wang Q, de Santiago I, Hopper JL, Tsimiklis H, et al. Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am J Hum Genet. 2013; 93: 1046-1060.

35. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK, Wang Q, Dicks E, Lee A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013; 45: 353–361.

36. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, Orr N, Rhie SK, Riboli E, Feigelson HS, Le Marchand L, Buring JE, Eccles D, et al. Genome-wide

association studies identify four ER negative-specific breast cancer risk loci. Nat Genet. 2013; 45: 392-398, 398e1-2.

37. Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Edwards SL, Pickett HA, Shen HC, Smart CE, Hillman KM, Mai PL, Lawrenson K, 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: 371-384.

38. Ahsan H, Halpern J, Kibriya MG, Pierce BL, Tong L, Gamazon E, McGuire V, Felberg A, Shi J, Jasmine F, Roy S, Brutus R, Argos M, et al. A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. Cancer Epidemiol Biomarkers Prev. 2014; 23: 658-669.

39. Milne RL, Burwinkel B, Michailidou K Arias-Perez JI, Zamora MP, Menéndez-Rodríguez P, Hardisson D, Mendiola M, González-Neira A, Pita G, Alonso MR, Dennis J, Wang Q, et al. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. Hum Mol Genet. 2014; 23: 6096-6111.

40. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, Maranian MJ, Bolla MK, Wang Q, Shah M, Perkins BJ, Czene K, Eriksson M, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015; 47: 373-380.

41. Orr N, Dudbridge F, Dryden N, Maguire S, Novo D, Perrakis E, Johnson N, Ghoussaini M, Hopper JL, Southey MC, Apicella C, Stone J, Schmidt MK, et al. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Hum Mol Genet. 2015; 24: 2966-2984.

42. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011; 43: 513–518.

43. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337: 1190-1195.

44. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow H, Lee MP. Allelic variation in gene expression is common in the human genome. Genome Res. 2003; 13: 1855–1862.

45. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. Science. 2002; 297: 1143.

46. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430: 743-747.

47. Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC, Gagné V, Dias J, Hoberman R, Montpetit A, et al. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet. 2009; 41: 1216-1222.

48. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015; 16: 197-212.

49. Li Q, Stram A, Chen C, Kar S, Gayther S, Pharoah P, Haiman C, Stranger B, Kraft P, Freedman ML. Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. Hum Mol Genet. 2014; 23: 5294-5302.

50. Li Q, Seo JH, Stranger B, McKenna A, Pe'er I, Laframboise T, Brown M, Tyekucheva S, Freedman ML. Integrative eQTL-based analyses the biology of breast cancer risk loci. Cell. 2013; 152: 633-641.

51. Takata, K, Reh, S, Tomida, J, Person, MD and Wood, RD. Human DNA helicase HELQ participates in DNA interstrand crosslink tolerance with ATR and RAD51 paralogs. Nat Commun. 2013; 4: 2338.

52. Adelman CA, Lolo RL, Birkbak NJ, Murina O, Matsuzaki K, Horejsi Z, Parmar K, Borel V, Skehel JM, Stamp G. HELQ promotes RAD51 paralogue-dependent repair to avert germ cell loss and tumorigenesis. Nature. 2013; 502: 381-384.

53. Stolk L, Perry JR, Chasman DI, He C, Mangino M, Sulem P, Barbalic M, Broer L, Byrne EM, Ernst F, Esko T, Franceschini N, Gudbjartsson DF, et al. Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. Nat Genet. 2012; 44: 260-268.

54. Day FR, Ruth KS, Thompson DJ, Lunetta KL, Pervjakova N, Chasman DI, Stolk L, Finucane HK, Sulem P, Bulik-Sullivan B, Esko T, Johnson AD, Elks CE, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat Genet. 2015; 47: 1294-1303.

55. McKay JD, Truong T, Gaborieau V, Chabrier A, Chuang SC, Byrnes G, Zaridze D, Shangina O, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Bucur A, et al. A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. PLoS Genet. 2011; 7: e1001333.

56. Babron MC, Kazma R, Gaborieau V, McKay J, Brennan P, Sarasin A, Benhamou S. Genetic variants in DNA repair pathways and risk of upper aerodigestive tract cancers: combined analysis of data from two genome-wide association studies in European populations. Carcinogenesis. 2014; 35: 1523-1527.

57. Pelttari LM, Kinnunen L, Kiiski JI, Khan S, Blomqvist C, Aittomäki K, Nevanlinna H. Screening of HELQ in breast and ovarian cancer families. Fam Cancer. 2016; 15: 19-23.

58. Cox DG, Simard J, Sinnett D, Hamdi Y, Soucy P, Ouimet M, Barjhoux L, Verny-Pierre C, McGuffog L, Healey S, Szabo C, Greene MH, Mai PL, et al. Common variants of the BRCA1 wild-type allele modify the risk of breast cancer in BRCA1 mutation carriers. Hum Mol Genet. 2011; 20: 4732-4747.

59. Maia AT, Antoniou AC, O'Reilly M, Samarajiwa S, Dunning M, Kartsonaki C, Chin SF, Curtis CN, McGuffog L, Domchek SM, EMBRACE, Easton DF, Peock S, et al. Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. Breast Cancer Res. 2012; 14: R63.

60. Lin W, Lin HD, Guo XY, Lin Y, Su FX, Jia WH, Tang LY, Zheng W, Long JR, Ren ZF. Allelic expression imbalance polymorphisms in susceptibility chromosome regions and the risk and survival of breast cancer. Mol Carcinog. 2016; Apr 29. doi: 10.1002/mc.22493. [Epub ahead of print].

61. Wang B, Matsuoka S, Ballif BA, Zhang D, Smogorzewska A, Giyi S, Elledge SJ. Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. Science. 2007; 316: 1194-1198.

62. Liu Z, Wu J, Yu X. CCDC98 targets BRCA1 to DNA damage sites. Nat Struct Mol Biol. 2007; 14: 716-720.

63. Kim H, Huang J, Chen J. CCDC98 is a BRCA1-BRCT domain-binding protein involved in the DNA damage response. Nat Struct Mol Biol. 2007; 14: 710-715.

64. Castillo, A, Paul, A, Sun, B, Huang, TH, Wang, Y, Yazinski, SA, Tyler, J, Li, L, You, MJ, Zou, L, Yao J, Wang B. The BRCA1-interacting protein Abraxas is required for genomic stability and tumor suppression. Cell Rep. 2014; 8: 807-817.

65. Solyom S, Aressy B, Pylkas K, Patterson-Fortin J, Hartikainen JM, Kallioniemi A, Kauppila S, Nikkilä J, Kosma VM, Mannermaa A, Greenberg RA, Winqvist R. Breast cancer-associated Abraxas mutation disrupts nuclear localization and DNA damage response functions. Sci Transl Med. 2012; 4: 122ra23.

66. Renault AL, Lesueur F, Coulombe Y, Gobeil S, Soucy P, Hamdi Y, Desjardins S, Le Calvez-Kelm F, Vallée M, Voegele C, The Breast Cancer Family Registry, Hopper JL, Andrulis IL, et al. ABRAXAS (FAM175A) and Breast Cancer Susceptibility: No Evidence of Association in the Breast Cancer Family Registry. PLoS One. 2016; 11: e0156820.

67. Snopok B, Yurchenko M, Szekely L, Klein G, Kashuba E. SPR-based immunocapture approach to creating an interfacial sensing architecture: Mapping of the MRS18-2 binding site on retinoblastoma protein. Anal Bioanal Chem. 2006; 386: 2063-2073.

68. Kashuba E, Yurchenko M, Yenamandra SP, Snopok B, Isaguliants M, Szekely L, Klein G. EBV-encoded EBNA-6 binds and targets MRS18-2 to the nucleus, resulting in the disruption of pRb-E2F1 complexes. Proc Natl Acad Sci USA. 2008; 105: 5489-5494.

69. Darekar SD, Mushtaq M, Gurrapu S, Kovalevska L, Drummond C, Petruchek M, Tirinato L, Di Fabrizio E, Carbone E, Kashuba E. Mitochondrial ribosomal protein S18-2 evokes chromosomal instability and transforms primary rat skin fibroblasts. Oncotarget. 2015; 6: 21016-21028.

70. Adoue V, Schiavi A, Light N, Almlöf JC, Lundmark P, Ge B, Kwan T, Caron M, Rönnblom L, Wang C, Chen SH, Goodall AH, Cambien F, et al. Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. Mol Syst Biol. 2014; 10: 754.

71. Almlöf JC, Lundmark P, Lundmark A, Ge B, Maouche S, Göring HH, Liljedahl U, Enström C, Brocheton J, Proust C. Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. PLoS One. 2012; 7: e52260.

72. Grundberg E, Adoue V, Kwan T, Ge B, Duan QL, Lam KC, Koka V, Kindmark A, Weiss ST, Tantisira K. Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. PLoS Genet. 2011; 7: e1001279.

73. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491: 56-65.

74. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156-2158.

75. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Maller J, Sklar P, de Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559-575.

76. Shin J-H, Blay S, McNeney B, Graham J. LD heatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. J Stat Soft. 2006; 16 Code Snippet 3.

77. Couch FJ, Wang X, McGuffog L, Lee A, Olswold C, Kuchenbaecker KB, Soucy P, Fredericksen Z, Barrowdale D, Dennis J, Gaudet MM, Dicks E, Kosel M, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet. 2013; 9: e1003212.

78. Gaudet MM, Kuchenbaecker KB, Vijai J, Klein RJ, Kirchhoff T, McGuffog L, Barrowdale D, Dunning AM, Lee A, Dennis J, Healey S, Dicks E, Soucy P, et al. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk PLoS Genet. 2013; 9 :e1003173.

79. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22: 1790-1797.

80. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research. 2012; 40: D930–D934 Database issue.

81. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155: 934-947.

82. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci USA. 2014; 111: E2191-E2199.

83. Beauparlant CJ, Lemacon A, Droit A. *ENCODExplorer: A compilation of ENCODE metadata*. R package version 1.4.3. 2015.

84. Joly Beauparlant C, Lamaze F, Deschenes A, Samb R, Lemaçon A, Belleau P, Bilodeau S, Droit A. *Metagene* profiles analyses reveal regulatory element's factor-specific recruitment patterns. PLOS Comput Biol. 2016 12:e1004751.

85. Haakensen VD, Lingjaerde OC, Lüders T, Riis M, Prat A, Troester MA, Holmen MM, Frantzen JO, Romundstad L, Navjord D, Bukholm IK, Johannesen TB, Perou CM, et al. Gene expression profiles of breast biopsies from healthy women identify a group with claudin-low features. BMC Med Genomics. 2011; 4: 77.

86. Quigley DA, Fiorito E, Nord S, Van Loo P, Alnæs GG, Fleischer T, Tost J, Moen Vollan HK, Tramm T, Overgaard J, Bukholm IR, Hurtado A, Balmain A, et al. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. Mol Oncol. 2014; 8: 273-284.

87. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490: 61-70.

88. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010; 26: 493-500.

89. Naume B, Zhao X, Synnestvedt M, Borgen E, Russnes HG, Lingjaerde OC, Strømberg M, Wiedswang G, Kvalheim G, Karesen R. Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. Mol Oncol. 2007; 1: 160-171.

90. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007; 23: 1294-1296.

Table 1: Associations with breast cancer risk for SNPs showing evidence of differential allelic expression (overall p <0.01)

SNP	Chr[a]	Position[b]	Alleles[c]	MAF[d]	OR[e] _ overall risk (95%CI)	*P*1df[f] _ overall risk	OR[e]_ER+ (95% CI)	*P*1df[f] _ ER+	OR[e]_ER- (95% CI)	*P*1df[f] _ ER-	Genes
rs697004	1	211842808	G/T	0.32	0.96 (0.94-0.98)	$3.67x10^{-04}$	0.96 (0.94-0.99)	$2.84x10^{-03}$	0.97 (0.93-1.01)	$1.46x10^{-01}$	*NEK2*
rs13447450	1	91965850	C/T	0.37	0.97 (0.95-0.99)	$1.16x10^{-03}$	0.96 (0.94-0.99)	$1.53x10^{-03}$	0.96 (0.92-1.00)	$3.61x10^{-02}$	*CDC7*
rs12125947	1	91990487	T/C	0.49	0.97 (0.95-0.99)	$1.59x10^{-03}$	0.97 (0.95-0.99)	$4.23x10^{-03}$	0.97 (0.93-1.01)	$9.47x10^{-02}$	*CDC7*
rs10490250	2	58509628	A/C	0.21	1.03 (1.01-1.06)	$7.51x10^{-03}$	1.03 (1.01-1.06)	$1.85x10^{-02}$	1.05 (1.01-1.10)	$4.29x10^{-02}$	*FANCL*
rs13099560	3	48204768	C/T	0.34	0.97 (0.95-0.99)	$3.99x10^{-03}$	0.96 (0.94-0.98)	$8.75x10^{-04}$	1.00 (0.96-1.04)	$9.24x10^{-01}$	*CDC25A*
rs11099601	4	84382763	A/G	0.50	1.05 (1.03-1.07)	**$5.62x10^{-06}$**	1.05 (1.03-1.08)	**$5.22x10^{-06}$**	1.07 (1.03-1.11)	$4.08x10^{-04}$	*HELQ, MRPS18C, FAM175A*
rs17355027	4	84388915	C/T	0.08	0.95 (0.92-0.98)	$4.46x10^{-03}$	0.95 (0.91-0.99)	$1.17x10^{-02}$	0.92 (0.86-0.98)	$1.55x10^{-02}$	*FAM175A*
rs2362974	5	36156654	C/T	0.12	0.96 (0.93-0.99)	$5.96x10^{-03}$	0.97 (0.93-1.00)	$5.45x10^{-02}$	0.99 (0.94-1.06)	$9.56x10^{-01}$	*SKP2*
rs733590	6	36645203	T/C	0.36	1.04 (1.03-1.06)	$1.77x10^{-04}$	1.03 (1.01-1.06)	$7.30x10^{-03}$	1.05 (1.01-1.09)	$1.99x10^{-02}$	*CDKN1A*
rs656040	11	65621057	C/T	0.33	1.05 (1.02-1.07)	**$1.52x10^{-05}$**	1.05 (1.03-1.07)	**$5.96x10^{-05}$**	1.03 (0.99-1.07)	$2.25x10^{-01}$	*SNX32, CFL1, MUS81*
rs570933	15	43824030	T/C	0.29	0.97 (0.95-0.99)	$7.18x10^{-03}$	0.97 (0.95-0.99)	$1.78x10^{-02}$	0.96 (0.92-1.00)	$2.89x10^{-02}$	*TP53BP1, MAP1A, HISPPD2A*
rs7234479	18	20599564	A/C	0.11	1.05 (1.02-1.08)	$1.59x10^{-03}$	1.05 (1.01-1.09)	$7.57x10^{-03}$	1.04 (0.98-1.10)	$1.52x10^{-01}$	*RBBP8*
rs738200	22	28792887	C/T	0.10	1.07 (1.03-1.10)	**$5.32x10^{-05}$**	1.09 (1.05-1.13)	**$7.21x10^{-06}$**	1.03 (0.97-1.09)	$3.87x10^{-01}$	*TTC28, CHEK2*

[a] Chromosome
[b] Build 37 position
[c] Major/minor allele, based on the forward strand and minor allele frequency in Europeans
[d] Mean minor allele frequency over all European controls in iCOGS
[e] Per-allele OR for the minor allele relative to the major allele
[f] One-degree-of-freedom *P*-value

SNPs highlighted in bold are those with associations for overall breast cancer risk reaching $p<10^{-4}$ (significance cut-off after Bonferroni correction)

Legends to Tables and Figures

Table 1. Associations with breast cancer risk for SNPs showing evidence of differential allelic expression (overall *p* <0.01).

Fig. 1. Regional plots of breast cancer risk association at 4q21. Regional plot of association result, recombination hotspots and LD for the 4q21: 84,132,874-84,631,193 loci. The index SNP rs11099601 is plotted as a blue triangle. Directly genotyped SNPs are represented as triangles and imputed SNPs (r^2>0.3, MAF>0.02) are represented as circles. The LD (r^2) for the index SNP with each SNP was computed based on European ancestry subjects included in the 1000 Genome Mar 2012 EUR. Pairwise r^2 values are plotted using a red scale, where white and red signify r^2=0 and 1, respectively. *P*-values were from the single-marker analysis based on logistic regression models after adjusted for age, study sites and the first six principal components plus one additional principal component for the LMBC in analyses of data from European descendants. SNPs are plotted according to their chromosomal position: physical locations are based on GRCh37/hg19. Gene annotation was based on the NCBI RefSeq genes from the UCSC Genome Browser.

Fig. 2. Functional annotation of the 4q21 locus (A) Functional annotations using data from the ENCODE and NIH Roadmap Epigenomics projects. From top to bottom, epigenetic signals evaluated included DNase clusters in MCF7 and HMEC cells, chromatin state segmentation by Hidden Markov Model (ChromHMM) in HMEC, breast myoepithelial cells (BMC) and Variant human mammary epithelial cells (vHMEC), where red represents an active promoter region, orange a strong enhancer and yellow a poised enhancer respectively (the detailed color scheme of chromatin states is described in the UCSC browser), histone modifications in MCF7, HMEC and

BMC cell lines ; and overlap between candidate variants and Max binding site in MCF7 cells. All tracks were generated by the UCSC genome browser (hg 19). (B) Long-range chromatin interactions. From top to bottom, ChIA-Pet interactions for PolII and CTCF in MCF7 cells and Hi-C interactions in HMEC cells. The ChIA-PET raw data available on GEO under the following accession (GSE63525.K56, GSE33664, GSE39495) were processed with the GenomicRanges package. (C) Maps of mammary cell super-enhancer locations as defined in Hnisz et al. are shown in HMEC cells. Predicted enhancer-promoter determined interactions in MCF7 and HMEC cells, as defined by the integrated method for predicting enhancer targets (IM-PET) are shown. (D) RNA-Seq data from MCF7 and HMEC cell lines. The value of the RNA-Seq analysis corresponds to the mean RPM value for *FAM175A*, *MRPS18C*, *HELQ*, *AGPAT9*, *HSPE* and *COQ2* from four HMEC and 19 MCF7 datasets, respectively. The annotation was obtained through the Bioconductor annotation package TxDb.Hsapiens.UCSC.hg19.knownGene. The tracks have been generated using ggplot2 and ggbio library in R.

Fig. 3. Boxplots representing differential expression of *HELQ (A)*, *MRPS18C (B)*, *FAM175A* (C) and *HPSE* (D) in breast tissues. Differential expression between normal breast and tumor tissue was determined by a Kruskal-Wallis rank sum test using TCGA breast cancer RNAseq data from primary tumor, metastasis and adjacent normal. Horizontal bars indicate mean expression levels.

Fig. 4. Boxplots representing expression levels of *HELQ (A)*, *MRPS18C (B)*, *FAM175A* (C) and *HPSE* (D) in the 5 molecular subtypes (PAM50 classifier) of breast primary tumors. Differential expression between normal breast and tumor tissue was determined by a Kruskal-Wallis rank sum test. Analysis was performed using TCGA breast cancer RNAseq data from five molecular subtypes of breast primary tumors : Luminal A (LumA), Luminal B (LumB), Human

epidermal growth factor receptor 2-enriched (Her2), Basal-like (Basal) and Normal-like (Normal). Horizontal bars indicate mean expression levels.

Fig. 5. Manhattan plots of association for the eQTL results at the 4q21 locus in normal breast and breast cancer tissue. *Y*-axis shows –log10(*P*-value) while *x*-axis shows physical position. Circles of various shades of blue represent breast cancer risk associations for all breast cancer tumors, ER+ and ER- tumors. Other colored circles represent eQTL results in the following datasets: normal breast (NB93, NB116) in various shades of green, breast carcinomas in pink (BC241) and red (BC765). Risk association results as well as eQTL results are for both imputed and genotyped SNPs for all datasets.

Fig. 6. Boxplots representing the most significant eQTL results for variant rs11099601 in normal breast tissue and breast tumor datasets. Box plots represent the expression levels of the indicated transcripts with respect to the rs11099601 genotypes. Expression levels are shown for A) *HELQ* in breast carcinoma BC241 dataset, B) *HELQ* in breast carcinoma BC765 dataset C) *HELQ* in breast carcinoma BC765 dataset normalized per isoform, D) *HELQ* in normal breast NB93 dataset E) *HELQ* in normal breast NB93 dataset normalized by gene isoform, F) *MRPS18C* in breast carcinoma BC765 dataset, G) *MRPS18C* in breast carcinoma BC765 dataset normalized per isoform, H) *FAM175A* in breast carcinoma BC765 dataset and I) *HSPE* in normal breast NB116 dataset. Horizontal bars indicate mean expression level per genotype. r^2 values indicate the percentage of variance in respective gene expression levels explained by rs11099601.

Supporting Information Captions

S1 Fig. Forest plots for the three most significant SNPs (overall *P*-value <10^{-4}). Squares indicate the estimated per-allele OR for the minor allele in Europeans. The horizontal lines indicate 95% confidence limits. The area of the square is inversely proportional to the variance of the estimate. The diamond indicates the estimated per-allele OR from the combined analysis.

S2 Fig. Differential allelic expression mapping of *FAM175A* and *MUS81* loci. (A) (C) The most significant *cis*-regulatory variants mapped by regression analysis in the primary monocyte population for *FAM175A* and *MUS81* are rs11099601 (*P*=5x10^{-22}) (A) and rs656040 (*P*=5.7x10^{-20}) respectively (C). Screenshot of the rs11099601 (B) and rs656040 (D) regions from the UCSC genome browser. Tracks display from top to bottom the *P*-values (-log10) of the allelic expression mapping in primary monocytes for each SNP, transcription factor binding (ENCODE ChIP-seq data) and average allelic expression across all individuals heterozygous for rs11099601 (B) and rs656040 (D).

S3 Fig. Distribution of the 5 molecular subtypes (PAM50 classifier) of breast primary tumors in the two breast cancer samples sets used for eQTL analysis – BC241 and BC765. The distribution of Luminal A, Luminal B, Human epidermal growth factor receptor 2-enriched (Her2), Basal-like and Normal-like subtypes is represented as the percentage of the total number of samples in each sample set.

S1 Table. List of selected genes and genetic variants associated with differential allelic expression.

S2 Table. Associations for the 313 genotyped SNPs with overall, ER-positive and ER-negative breast cancer risk.

S3 Table. Associations for imputed and genotyped SNPs in the 4q21 locus (4q21: 84,132,874-84,631,193) for overall, ER-positive and ER-negative breast cancer risk.

S4 Table. Regulome DB analysis of SNPs in the 4q21 locus (4q21: 84,132,874-84,631,193) with r^2>0.8 with top associated SNP rs11099601. The scoring scheme refers to the following available datatypes: **1a** = eQTL + transcription factor (TF) binding + matched TF motif + matched DNase Footprint + DNase peak; **1b** = eQTL + TF binding + any motif + DNase Footprint + DNase peak; **1c** = eQTL + TF binding + matched TF motif + DNase peak; **1d** = eQTL + TF binding + any motif + DNase peak; **1e** = eQTL + TF binding + matched TF motif; **1f** = eQTL + TF binding / DNase peak; **2a** = TF binding + matched TF motif + matched DNase Footprint + DNase peak; **2b** = TF binding + any motif + DNase Footprint + DNase peak; **2c** = TF binding + matched TF motif + DNase peak; **3a** = TF binding + any motif + DNase peak; **3b** = TF binding + matched TF motif; **4** = TF binding + DNase peak; **5** = TF binding or DNase peak; **6** = other.

S5 Table. Description of the BCAC studies with subjects of European origin contributing to iCOGs.

S6 Table. Data sources for *in silico* analyses of the 4q21 breast cancer susceptibility loci

Figure 1

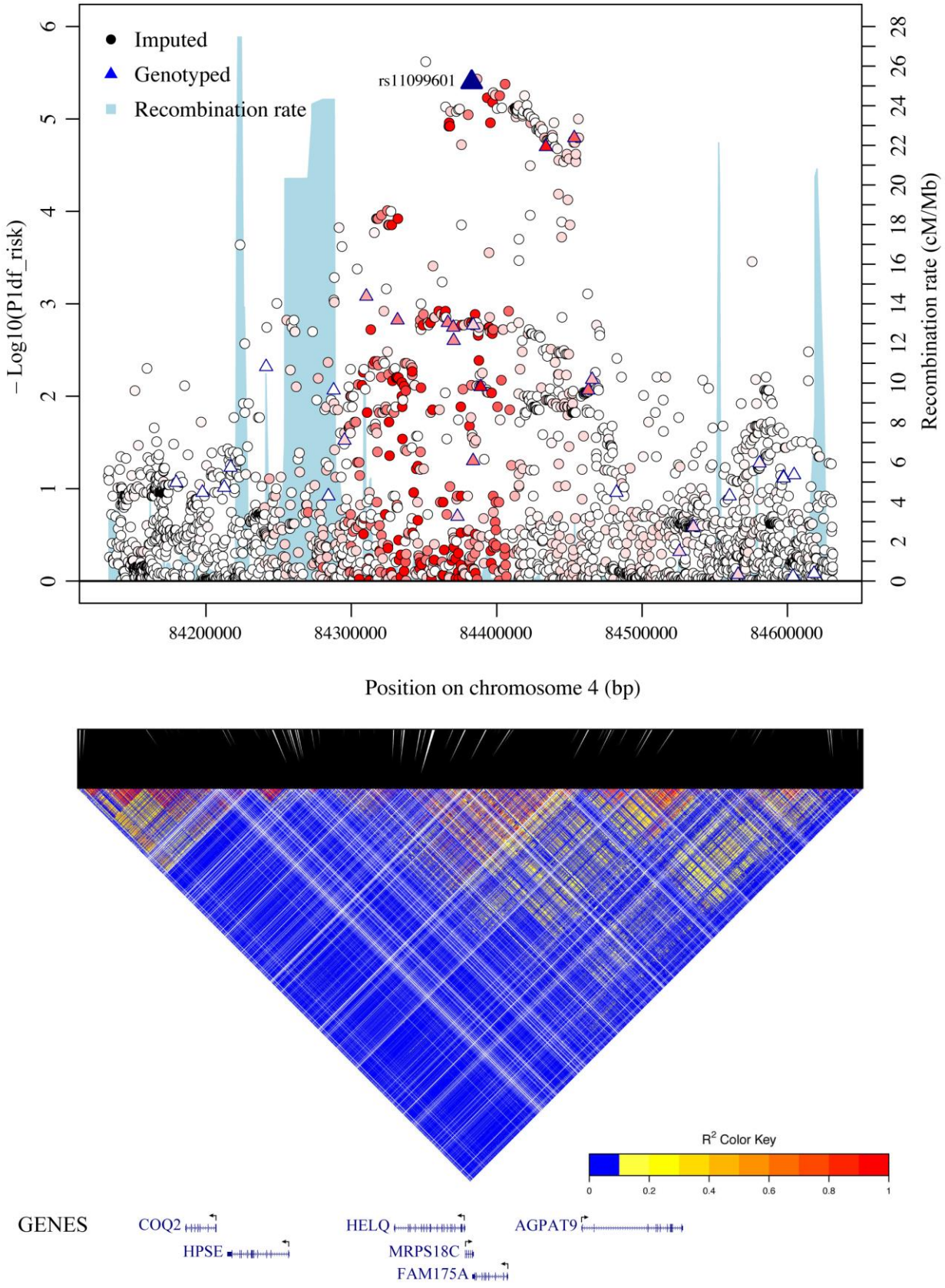

Imputed
Genotyped
Recombination rate
rs11099601
− Log10(P1df_risk)
Recombination rate (cM/Mb)
84200000
84300000
84400000
84500000
84600000
Position on chromosome 4 (bp)
R2 Color Key
GENES
COQ2
HPSE
HELQ
MRPS18C
FAM175A
AGPAT9

Figure 2

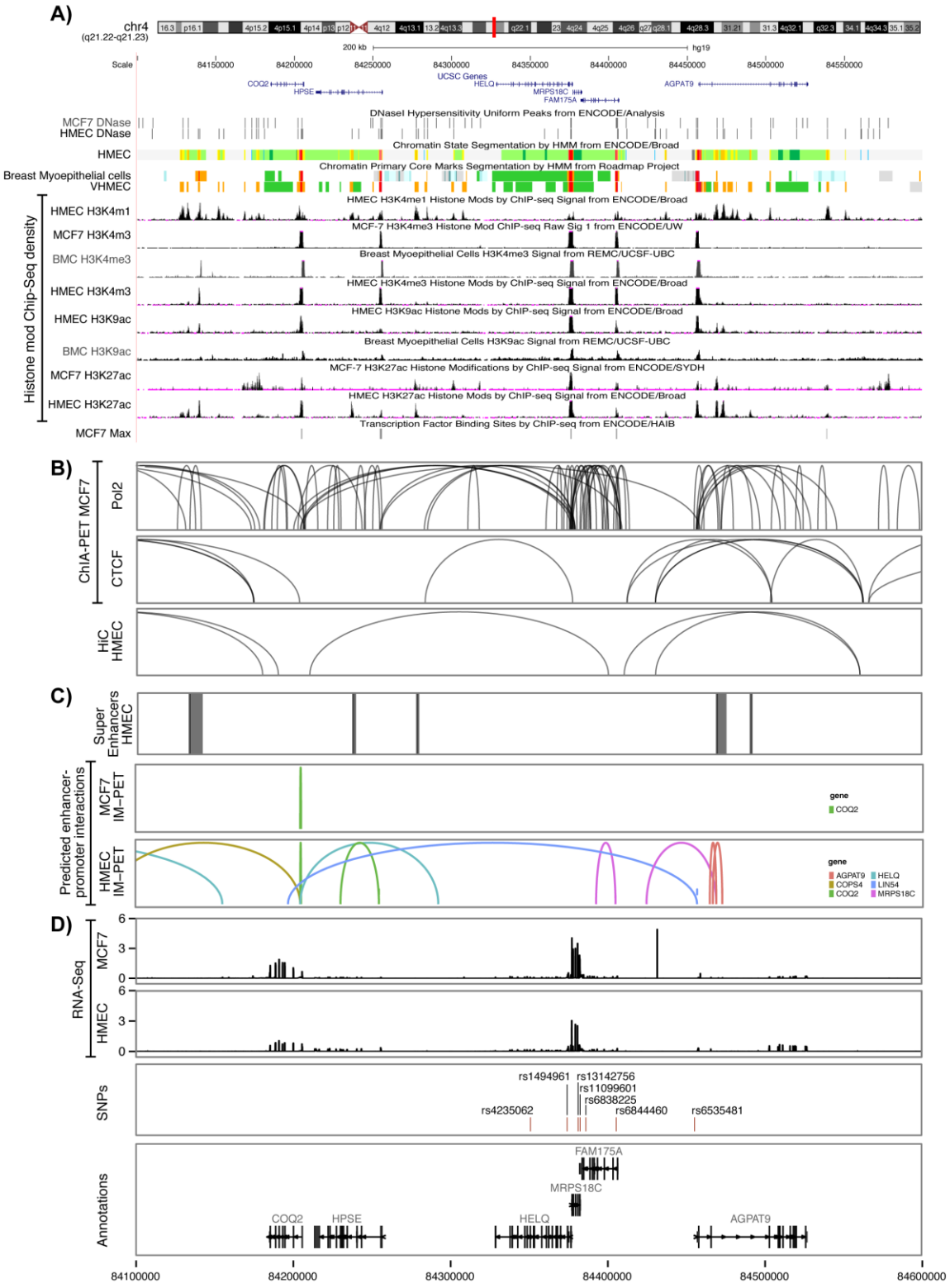

A)
chr4
(q21.22-q21.23)
200 kb
hg19
Scale
84150000
84200000
84250000
84300000
84350000
84400000
84450000
84500000
84550000
UCSC Genes
COQ2
HPSE
HELQ
MRPS18C
FAM175A
AGPAT9
DNaseI Hypersensitivity Uniform Peaks from ENCODE/Analysis
MCF7 DNase
HMEC DNase
Chromatin State Segmentation by HMM from ENCODE/Broad
HMEC
Chromatin Primary Core Marks Segmentation by HMM from Roadmap Project
Breast Myoepithelial cells
VHMEC
Histone mod Chip-Seq density
HMEC H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/Broad
HMEC H3K4m1
MCF-7 H3K4me3 Histone Mod ChIP-seq Raw Sig 1 from ENCODE/UW
MCF7 H3K4m3
Breast Myoepithelial Cells H3K4me3 Signal from REMC/UCSF-UBC
BMC H3K4me3
HMEC H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/Broad
HMEC H3K4m3
HMEC H3K9ac Histone Mods by ChIP-seq Signal from ENCODE/Broad
HMEC H3K9ac
Breast Myoepithelial Cells H3K9ac Signal from REMC/UCSF-UBC
BMC H3K9ac
MCF-7 H3K27ac Histone Modifications by ChIP-seq Signal from ENCODE/SYDH
MCF7 H3K27ac
HMEC H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/Broad
HMEC H3K27ac
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
MCF7 Max
B)
ChIA-PET MCF7
Pol2
CTCF
HiC
HMEC
C)
Super
Enhancers
HMEC
Predicted enhancer-
promoter interactions
MCF7
IM–PET
HMEC
IM–PET
gene
COQ2
gene
AGPAT9
HELQ
COPS4
LIN54
COQ2
MRPS18C
D)
RNA-Seq
MCF7
HMEC
6
3
0
SNPs
rs1494961
rs13142756
rs11099601
rs6838225
rs4235062
rs6844460
rs6535481
Annotations
FAM175A
MRPS18C
COQ2
HPSE
HELQ
AGPAT9
84100000
84200000
84300000
84400000
84500000
84600000

Figure 3

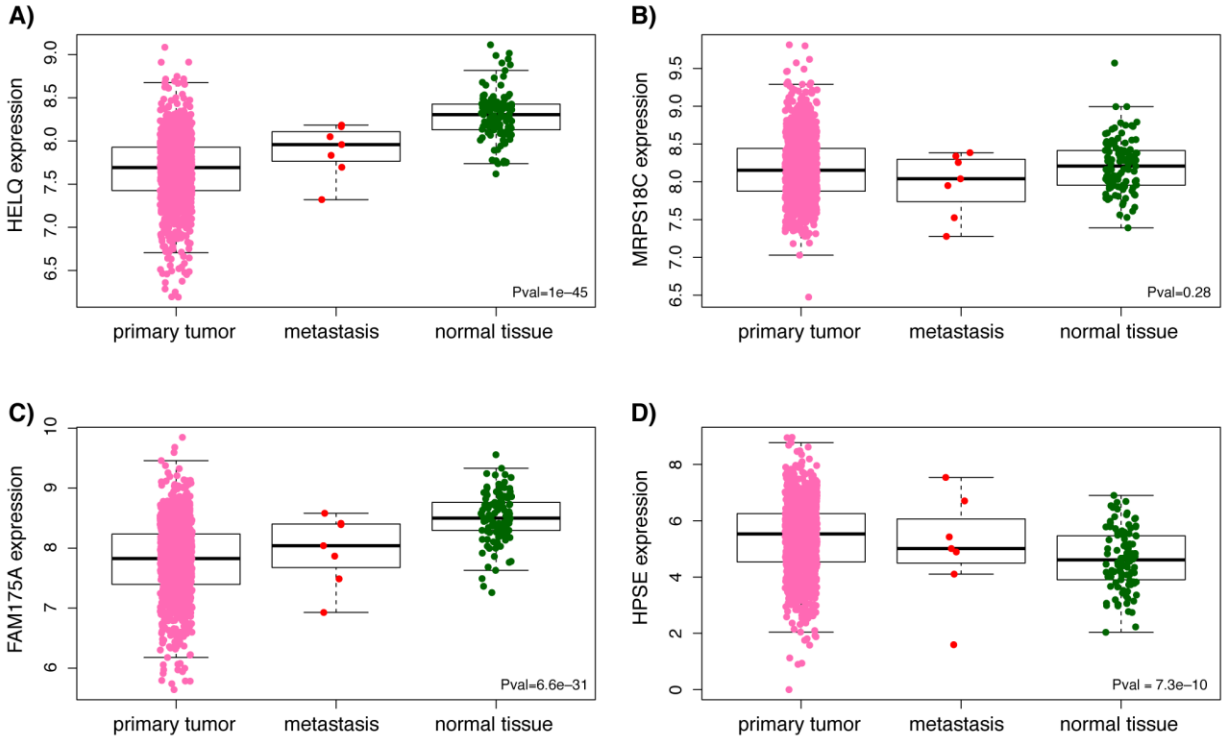

A)
HELQ expression
6.5
7.0
7.5
8.0
8.5
9.0
Pval=1e−45
primary tumor
metastasis
normal tissue
B)
MRPS18C expression
6.5
7.0
7.5
8.0
8.5
9.0
9.5
Pval=0.28
primary tumor
metastasis
normal tissue
C)
FAM175A expression
6
7
8
9
10
Pval=6.6e−31
primary tumor
metastasis
normal tissue
D)
HPSE expression
0
2
4
6
8
Pval = 7.3e−10
primary tumor
metastasis
normal tissue

Figure 4

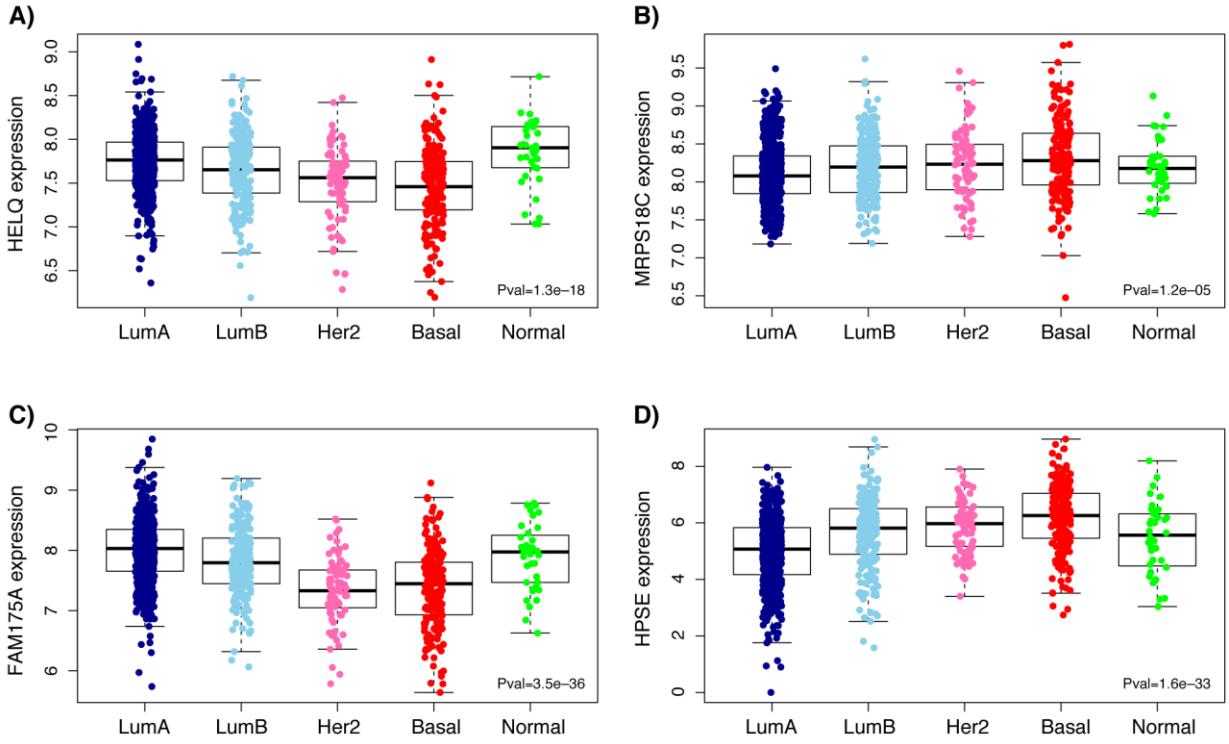

A)
HELQ expression
6.5 7.0 7.5 8.0 8.5 9.0
Pval=1.3e−18
LumA LumB Her2 Basal Normal
B)
MRPS18C expression
6.5 7.0 7.5 8.0 8.5 9.0 9.5
Pval=1.2e−05
LumA LumB Her2 Basal Normal
C)
FAM175A expression
6 7 8 9 10
Pval=3.5e−36
LumA LumB Her2 Basal Normal
D)
HPSE expression
0 2 4 6 8
Pval=1.6e−33
LumA LumB Her2 Basal Normal

Figure 5

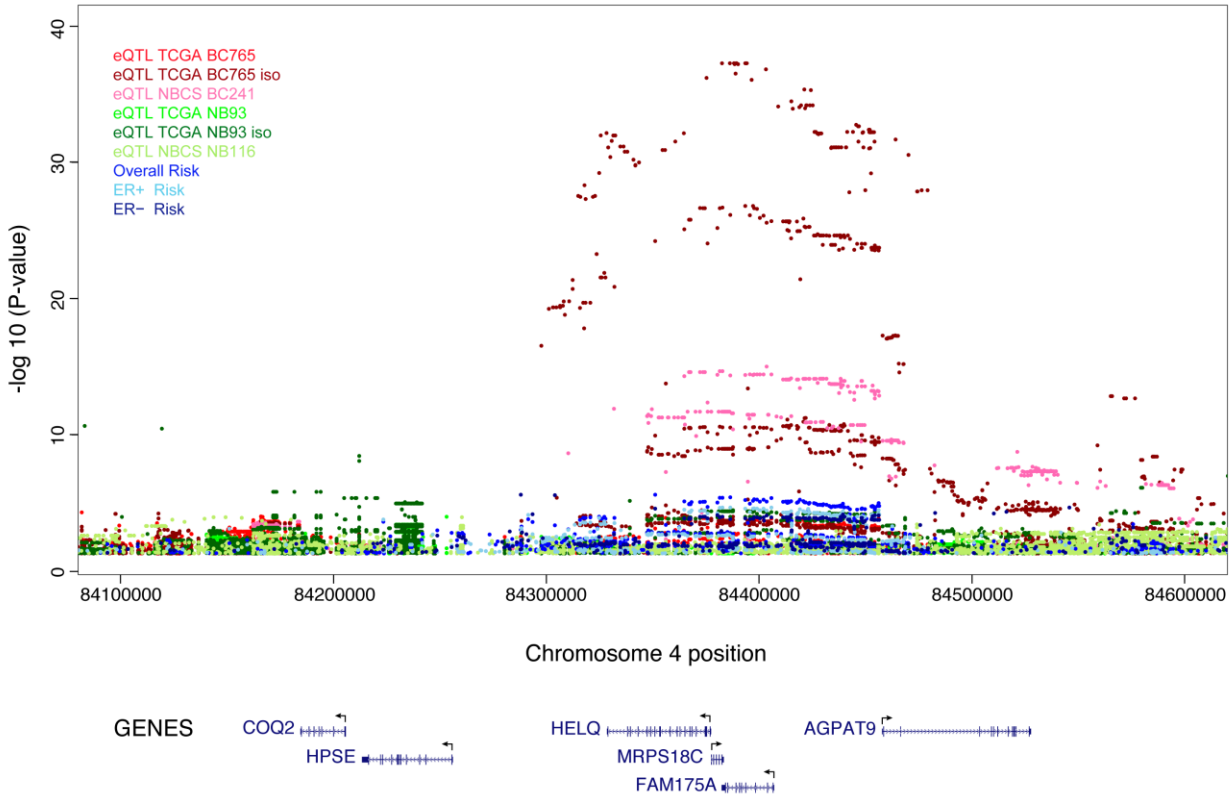
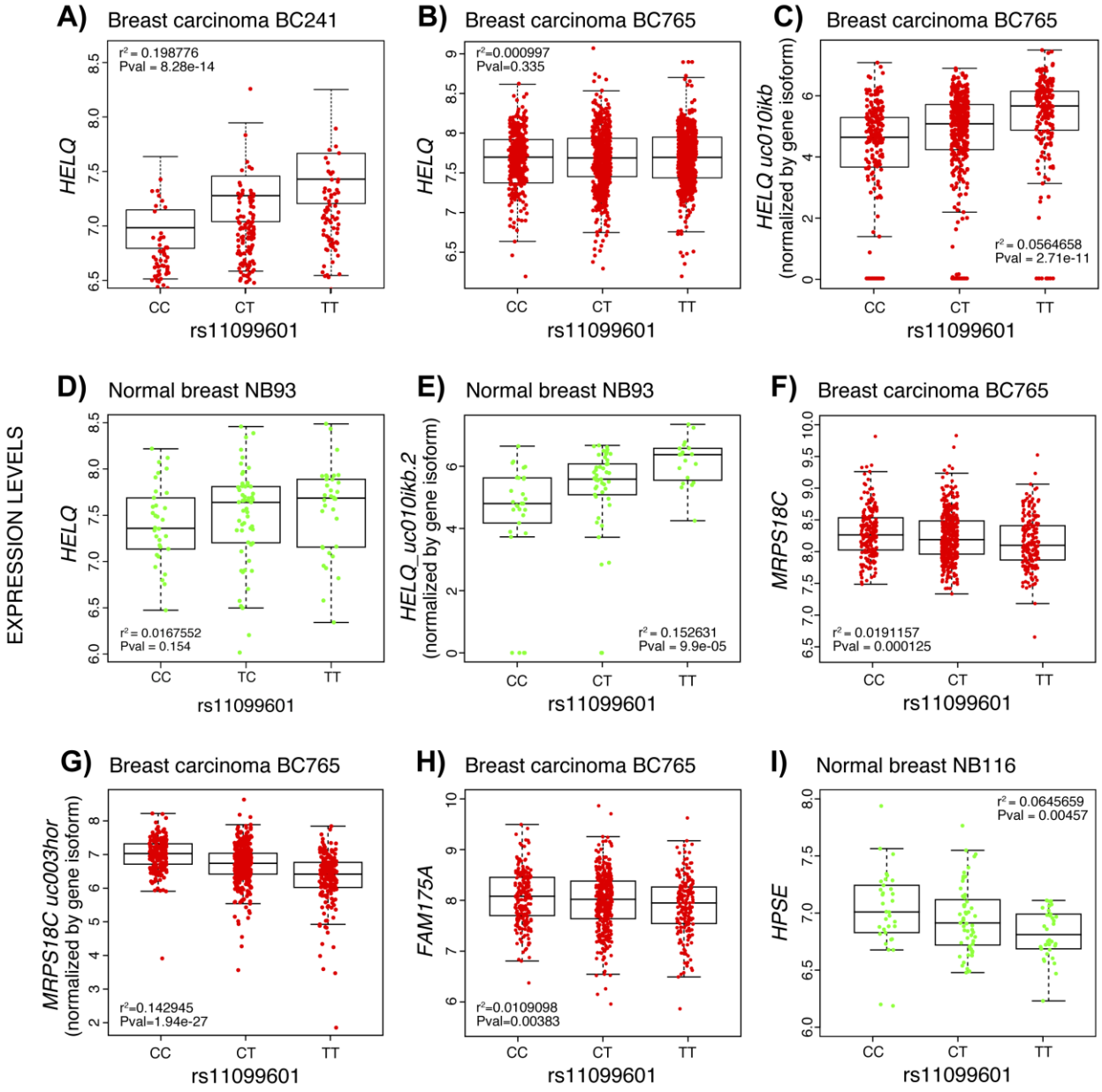

eQTL TCGA BC765
eQTL TCGA BC765 iso
eQTL NBCS BC241
eQTL TCGA NB93
eQTL TCGA NB93 iso
eQTL NBCS NB116
Overall Risk
ER+ Risk
ER− Risk
-log 10 (P-value)
0
10
20
30
40
84100000
84200000
84300000
84400000
84500000
84600000
Chromosome 4 position
GENES
COQ2
HPSE
HELQ
MRPS18C
FAM175A
AGPAT9

Figure 6


A) Breast carcinoma BC241
r² = 0.198776
Pval = 8.28e-14
HELQ
CC CT TT
rs11099601
B) Breast carcinoma BC765
r²=0.000997
Pval=0.335
HELQ
CC CT TT
rs11099601
C) Breast carcinoma BC765
HELQ uc010ikb
(normalized by gene isoform)
r² = 0.0564658
Pval = 2.71e-11
CC CT TT
rs11099601
EXPRESSION LEVELS
D) Normal breast NB93
r² = 0.0167552
Pval = 0.154
HELQ
CC TC TT
rs11099601
E) Normal breast NB93
HELQ_uc010ikb.2
(normalized by gene isoform)
r² = 0.152631
Pval = 9.9e-05
CC CT TT
rs11099601
F) Breast carcinoma BC765
MRPS18C
r² = 0.0191157
Pval = 0.000125
CC CT TT
rs11099601
G) Breast carcinoma BC765
MRPS18C uc003hor
(normalized by gene isoform)
r²=0.142945
Pval=1.94e-27
CC CT TT
rs11099601
H) Breast carcinoma BC765
FAM175A
r²=0.0109098
Pval=0.00383
CC CT TT
rs11099601
I) Normal breast NB116
r² = 0.0645659
Pval = 0.00457
HPSE
CC CT TT
rs11099601