Germline variants and breast cancer survival in patients with distant metastases at primary breast cancer diagnosis

Maria Escala-García1, Sander Canisius1,2, Renske Keeman1, Jonathan Beesley3, Hoda Anton-Culver4, Volker Arndt5, Annelie Augustinsson6, Heiko Becher7, Matthias W. Beckmann8, Sabine Behrens9, Marina Bermisheva10, Stig E. Bojesen11,12,13,14, Manjeet K. Bolla15, Hermann Brenner9,16,17, Federico Canzian18, Jose E. Castelao19, Jenny Chang-Claude6,20, Stephen J. Chanock21, Fergus J. Couch22, Kamila Czene23, Mary B. Daly24, Joe Dennis15, Peter Devilee25,26, Thilo Dörk15, Alison M. Dunning28, Douglas F. Easton25,28, Arif B. Eker31, A. Heather Eliassen30,31, Peter A. Fasching8,32, Henrik Flyer33, Manuela Gago-Dominguez14,35, Montserrat García-Closas21, José A. García-Sáenz36, Jürgen Geisler37, Graham G. Giles38,39,40, Mervi Grip41, Melanie Hahnen42,43,44, Eric Hahnen45,46, Christopher A. Haiman47, Niclas Håkansson48, Per Hall23,49, Ute Hamann50, Jaana M. Hartikainen51,52, Bernadette A. M. Heemskerk-Gerritsen53, Antoinette Hollestelle53, Reiner Hoppe54,55, John L. Hopper56, David J. Hunter51,56, William Jacot57, Anna Jakubowska58,59, Esther M. John60,61, Audrey Y. Jung9, Rudolf Kaaks9, Elza Khusnutdinova10,62, Linetta B. Koppert63, Peter Kraft31,64, Vessela N. Kristensen65, Allison W. Kuri60,61, Dietrich Lambrechts76,67, Loic Le Marchand68, Annika Lindblom69,70, Robert N. Luben71,72, Jan Lubinski39, Arto Manne10,45,46, Mehdi Manochehri50, Sara Margolin49,74, Dimitrios Mavroudis75, Tanja A. Muren76, Heli Nevanlinna76, Andrew F. Olshan77, Håkan Olsson6, Tjoung-Won Park-Simon77, Alpa V. Patel78, Paolo Peterlongo79, Paul D. Pharoah15,28, Kevin Punie69, Paolo Radice81, Gad Rennert82, Hedy S. Rennert82, Atocha Romero83, Rebecca Roylance84, Thomas Rüdiger85, Matthias Ruebner8, Emmanouil Saloustros96, Elinor J. Sawyer87, Rita K. Schmutzler55,46,88, Minouk J. Schoemaker89, Christopher Scott90, Melissa C. Southey38,40,91, Harald Surawy42,43, Anthony J. Swedlow89,92, Rulla M. Tamim31,93, Lauren R. Teras79, Emilie Thomas94, Ian Tomlinson95,96, Melissa A. Troester77, Celine M. Vachon97, Qin Wang98, Robert Winqvist98,99, Alicja Wolk48,100, Argyrios Zogas4, kConFab/AOCS Investigators101, Kyriaki Michailidou15,102,103, Georgia Chenevix-Trench3, Thomas Bachet60 & Marjanka K. Schmidt1,105.

Breast cancer metastasis accounts for most of the deaths from breast cancer. Identification of germline variants associated with survival in aggressive types of breast cancer may inform understanding of breast cancer progression and assist treatment. In this analysis, we studied the associations between germline variants and breast cancer survival for patients with distant metastases at primary breast cancer diagnosis. We used data from the Breast Cancer Association Consortium (BCAC) including 1062 women of European ancestry with metastatic breast cancer, 606 of whom died of breast cancer. We identified two germline variants on chromosome 1, rs138569520 and rs146023652, significantly associated with breast cancer-specific survival (P = 3.19 × 10−8 and 4.42 × 10−8). In silico analysis suggested a potential regulatory effect of the variants on the nearby target genes SDE2 and H3F3A. However, the variants showed no evidence of association in a smaller replication dataset. The validation dataset was obtained from the SNPs to Risk of Metastasis (StoRM) study and included 293 patients with metastatic primary breast cancer at diagnosis. Ultimately, larger replication studies are needed to confirm the identified associations.
Breast cancer is the most common female cancer in the Western world and one of the most common causes of cancer death in women globally. Early detection and better treatments have helped to reduce breast cancer mortality in recent decades. Yet, when breast cancer metastasizes to distant sites, prognosis continues to be poor and for most cases treatment is only palliative. Metastases in breast cancer can remain undetectable for many years after initial diagnosis, leading to incurable lesions. Approximately 15% of patients with breast cancer will develop distant metastases within 3 years after diagnosis of the primary tumor. Therefore, it is important to have the tools able to detect breast cancer metastases at earlier stages, in order to better manage and predict breast cancer progression. Prognostication models could benefit from the inclusion of germline genetic biomarkers that are capable of predicting tumor recurrence, second tumors or prognosis of second tumors. However, so far, it has been difficult to identify individual common germline variants associated with primary breast cancer survival due to the small effect size these variants are likely to have. Likewise, evidence as to whether or not germline variants can increase the probability of metastatic progression is currently limited to a few studies. For example, a candidate gene study identified common single nucleotide polymorphisms (SNPs) located within *SIPA1* that were associated with metastasis and poor breast cancer prognosis. Other studies have identified other metastasis susceptibility genes such as *RRP1b*. Germline variants could specifically provide metastatic predisposition by affecting treatment response or promoting tumor initiating events and providing new metastatic functions to tumor cells.

The aim of this study was to identify associations between common germline variants and breast cancer-specific survival in patients with metastasis at primary breast cancer diagnosis. We hypothesized that germline variants might predispose to poorer survival after breast cancer metastasis, and that analyzing a set of patients with similar stage of the disease might help identify variants that do not show evidence of association in larger but more heterogeneous datasets.

**Results**

We used data from the Breast Cancer Association Consortium (BCAC): the dataset comprised data from 50 studies from which follow-up information for women diagnosed with distant metastases at primary breast cancer diagnosis was available. The results were based on the meta-analysis of two genome-wide SNP arrays (iCOGS and OncoArray) (see “Methods”). We analyzed variants that had a minor allele frequency (MAF) > 0.01 and an imputation quality r^2 > 0.7 for at least one of the two arrays. Details about the individual studies, the genotyping
array used and number of patients included are given in Supplementary Table 1. We analyzed the genotypes and clinico-pathological data of a total of 1062 breast cancer patients, 606 of whom died of breast cancer within 15 years of follow-up. Of these, 721 of the patients had estrogen receptor (ER)-positive disease (388 deaths) and 227 had ER-negative disease (148 deaths). All patients were women of European descent. The patients were diagnosed from 1979 to 2014 (median: 2004) and aged 26–92 (median: 60) years.

Manhattan plots showing the association between germline variants and breast cancer-specific survival of all, ER-positive and ER-negative metastasized breast cancers are shown in Fig. 1. We identified two genome-wide significant (P < 5 × 10−8) variants (SNPs: rs138569520 and rs146023652) on chromosome 1 associated with breast cancer-specific survival for all metastasized breast cancers (Table 1, Supplementary Table 2). The two variants were part of a set of six highly correlated SNPs (Table 1, r2 > 0.88) based on European subjects in phase 3 of the 1000 Genomes Project44. No variant reached genome-wide significance for ER-positive or for ER-negative breast cancer tumors alone (Supplementary Tables 3 and 4).

The variant with the strongest association was the SNP rs138569520 (HR = 3.67, 95% CI 1.86–7.23 and P = 3.19 × 10−8). The HR estimates for rs138569520 in the ER-positive (HR = 3.38, 95% CI 1.48–7.70 and P = 4.37 × 10−4) and ER-negative (HR = 2.76, 95% CI 1.16–6.64 and P = 8.70 × 10−3) were similar (P = 0.97 for difference).

Several genes (SDE2, LEFTY2, PYCR2 and H3F3A) were located within 100 kb of the most significant SNP rs138569520. We interrogated functional genomic data including annotations of enhancers, promoters and transcription factor binding sites and found evidence consistent with gene regulation in the regions containing the associated variants (Fig. 2). Hi-C analysis in HMEC cells45 showed that the lead variant rs138569520 is located in a genomic region interacting with the promoter region of H3F3A. SNPs rs146023652 and rs114512448 overlapped with transcription factor (TF) binding sites which might reflect the active transcription of SDE2. ChIP-seq signals from primary breast sub-populations46 also showed potential regulatory regions containing rs114512448. ChIA-PET analysis in MCF-7 cells from ENCODE17, detected an interaction between rs114512448 and the PYCR2 gene. Finally, ChIA-PET also detected an interaction between rs72757046 and SDE2 and H3F3A.

Using KMplotter (kmplot.com/analysis)18, we tested the association of the mRNA tumor expression of SDE2 and H3F3A, the genes in closest proximity to rs138569520, with overall survival in grade 3 breast tumors (to select the most aggressive subtype; selection for stage 4 was not available). Low mRNA expression levels of SDE2 gene were significantly associated (P = 0.01) with poorer breast cancer survival (Fig. 3a), while, in contrast, high expression of H3F3A was associated with lower survival (P = 6.7 × 10−5) (Fig. 3b). These associations were not statistically significant, neither for grade 1 or for grade 2 disease (P > 0.21).

Lastly, we aimed to evaluate the significance of the two genome-wide significant SNPs using an independent set of 293 breast cancer patients with metastatic primary breast cancer at diagnosis from the SNPs to Risk of Metastasis (StoRM) study47. All patients were diagnosed in France from March 2012 to May 2014, aged 18 years or older (median: 59 years) and followed up to July 2017. A total of 293 patients were available for the validation study, 239 of whom had events, defined as progression and/or death occurring during follow-up. Both SNPs had good imputation quality (r2 ~ 0.7) and similar MAFs to those in the BCAC dataset (~ 2%). However, neither of the two SNPs replicated in the survival analysis with the StoRM dataset (Table 2): rs138569520 (HR = 1.49, 95% CI 0.46–3.37, P = 0.66) and rs146023652 (HR = 1.25, 95% CI 0.46–3.37, P = 0.66). Although the HR estimates in the StoRM validation dataset were smaller than those from the BCAC analyses (HR = 3.67 and 3.64), the confidence limits overlapped.

Because the BCAC dataset also included prevalent cases (n = 466), we repeated the analysis with incident cases (n = 596) to match the study design in StoRM more closely. The HR estimates were similar to those for...
In this analysis of breast cancer patients with metastatic primary breast cancer at diagnosis, involving 1062 patients with 606 breast cancer-specific deaths, we identified two variants on chromosome 1 (rs138569520 and rs146023652) associated with survival, at genome-wide levels of statistical significance. The most significant association was for the SNP rs138569520 (P = 3.19 × 10−8). The HR estimates were similar in patients with ER-positive and ER-negative disease.

Two genes, SDE2 and H3F3A, were in closest proximity of rs138569520. Both genes have been previously associated with oncogenic processes relevant for metastatic progression: the SDE2 gene ("silencing defective 2") is known to be involved in DNA replication, telomere maintenance and cell cycle control. The functional roles of SDE2 have been studied in a proteome dynamics analysis in prostate cancer cells; the results suggested that alterations of the gene might diminish the error-prone DNA repair pathway activation and promote missense mutations. The gene H3F3A encodes for histone H3.3, and mutations in this protein have been linked to multiple cancer processes and breast invasive ductal carcinoma. Additionally, the differential expression of these two genes was significantly associated with survival in grade 3 tumors based on KMplotter. Previous studies have also linked the expression of these genes to oncogenic processes. For example, downregulation of SDE2 was associated with mutation disease phenotype as well as poorer mortality outcomes. Likewise, overexpression of H3F3A was associated with lung cancer progression and promotion of lung cancer cell migration by activation of metastasis-related genes. Unfortunately, in KMplotter it was not possible to specifically select stage 4 tumors, which limits the interpretation of our findings. Future studies are needed in order to corroborate the association of SDE2 and H3F3A expression with survival in this group of patients.

Additionally, there was predicted genomic activity in the locus based on the intersection of multiple genomic regulatory features in breast tissue. Although the SNPs appeared to cluster around SDE2, there was also in-silico evidence for two other potential target genes at this locus (H3F3A and PYCR2). PYCR2 encodes for a mitochondrial protein involved in proline biosynthesis. While little is known about this proline form, studies for

Table 1. Results for the six correlated variants associated with breast cancer-specific survival for patients with metastatic primary breast cancer at diagnosis. Genomic positions are based on the hg19 genome build. ALT alternate, REF reference, EAF effect allele frequency, HR hazard rate, LCL lower control limit, UCL upper control limit, $r^2$ imputation quality.

| SNP             | Chr | Position     | Ref | Alt | EAF | $r^2$ | HR  | LCL  | UCL  | P value   |
|-----------------|-----|--------------|-----|-----|-----|------|-----|------|------|-----------|
| rs138569520     | 1   | 22613175     | T   | C   | 0.02| 0.87 | 3.67| 1.86| 7.23| 3.19 × 10−8 |
| rs146023652     | 1   | 226158826    | C   | T   | 0.02| 0.86 | 3.64| 1.84| 7.19| 4.42 × 10−8 |
| rs114512448     | 1   | 226173980    | G   | A   | 0.02| 0.86 | 3.53| 1.78| 6.95| 6.57 × 10−8 |
| rs143653255     | 1   | 226157179    | T   | C   | 0.02| 0.86 | 3.26| 1.68| 6.34| 1.53 × 10−7 |
| rs115086585     | 1   | 226154721    | C   | T   | 0.02| 0.85 | 3.21| 1.64| 6.25| 2.93 × 10−7 |
| rs72757046      | 1   | 226235714    | G   | C   | 0.02| 0.84 | 3.62| 1.78| 7.37| 6.02 × 10−7 |

Figure 2. Functional annotation of the six highly correlated SNPs: rs138569520, rs146023652, rs114512448, rs143653255, rs115086585 and rs72757046. TF transcription factor.

Discussion

In this analysis of breast cancer patients with metastatic primary breast cancer at diagnosis, involving 1062 patients with 606 breast cancer-specific deaths, we identified two variants on chromosome 1 (rs138569520 and rs146023652) associated with survival, at genome-wide levels of statistical significance. The most significant association was for the SNP rs138569520 (P = 3.19 × 10−8). The HR estimates were similar in patients with ER-positive and ER-negative disease.
the close family member \(PYCR1\) have found that higher levels of mRNA were associated with reduced survival from breast cancer patients\(^{26}\). To support further our hypothesis that the two genome-wide significant SNPs (rs138569520 and rs146023652) were specific for survival in patients with metastatic disease, we confirmed that there were no associations (HR = 1.04, \(P = 0.58\), MAF = 0.02 and HR = 1.03, \(P = 0.60\), MAF = 0.02 respectively) with breast cancer-specific survival in the most recent BCAC dataset for all invasive early (stages I–III) breast cancers (OncoArray and iCOGS, \(n = 86,627\))\(^{27}\).

On the other hand, the two genome-wide significant variants, rs138569520 and rs146023652, were not replicated (\(P = 0.34\) and \(P = 0.66\), respectively) using an independent dataset of patients with metastatic primary breast cancer at diagnosis (\(n = 293\)). The imputation quality and the minor allele frequency of the SNPs in the replication cohort were comparable to those in the BCAC analyses (MAF = 2% and \(r^2 > 7\%\)), therefore the negative result could not be attributed to those factors. Age of the patients could also not explain the difference since both datasets had comparable median ages at diagnosis, 60 years for BCAC and 59 years for StoRM. On the other hand, it is important to state that there were several factors that varied between the datasets. First, the sample size differed considerably between BCAC (\(n = 1062\)) and the StoRM study (\(n = 293\)), the latter having a relatively small sample size which limits the power to detect associations. Total follow-up time also varied: for the BCAC dataset, patients were followed for a maximum of 15 years, while for the StoRM study the follow-up ended at 5 years. However, the results from the complementary analysis using the BCAC dataset and 5-year follow-up were comparable to the initial 15 years follow-up results. This finding suggests that the disparity in estimates between the two analyses is not due to shorter follow-up. There were several other differences between the main BCAC dataset and the StoRM cohort used for validation. For example, the BCAC dataset included multiple studies from several countries while the StoRM cohort included solely patients from France. Moreover, StoRM was a recent cohort with the earliest reported diagnosis starting in 2012. On the other hand, in BCAC, the year of patients’ diagnosis ranged between 1979 and 2014 and included prevalent cases. While the analysis in BCAC using exclusively incident cases gave comparable estimates to the main analysis, the difference in the years of diagnosis could be related to differences in treatment strategies that were not considered in the current analysis. The lack of information about detailed treatment is a potential weakness of the current analysis and validation. Treatment strategy, together with characteristics of the tumor, will also influence the final prognosis of metastatic

Figure 3. Kaplan–Meier overall survival plot for high versus low expression level of the genes (A) \(SDE2\) (\(n = 204\)) and (B) \(H3F3A\) (\(n = 503\)) restricted to patients with a grade 3 tumor and 15 years of follow-up. The differential expression analysis was performed in KMplotter.

Table 2. Results for the validation of the two genome-wide significant variants in an independent dataset of breast cancer patients with metastatic primary breast cancer at diagnosis. ALT alternate, REF reference, EAF effect allele frequency, HR hazard rate, LCL lower control limit, UCL upper control limit, \(r^2\) imputation quality.

| SNP       | Chr | Position | Ref | Alt | EAF | \(r^2\) | HR   | LCL  | UCL  | \(P\) value |
|-----------|-----|----------|-----|-----|-----|--------|------|------|------|-------------|
| rs138569520 | 1   | 226193175 | T   | C   | 0.02 | 0.69   | 1.49 | 0.60 | 3.71 | 0.34         |
| rs146023652 | 1   | 226158826 | C   | T   | 0.02 | 0.79   | 1.25 | 0.46 | 3.37 | 0.66         |
breast cancer\textsuperscript{28}. It is important to note that the associations observed in the BCAC study may be false positives, and that further large replication studies will be required to confirm or refute the associations.

In conclusion, this analysis of BCAC patients with metastatic primary breast cancer at diagnosis from the BCAC dataset identified a new region in chromosome 1 associated with breast cancer-specific survival. The region includes six highly correlated SNPs that are predicted to be in an active region of the genome based on in-silico evidence from breast cancer tissues and that are located in close proximity to genes involved in oncogenic processes. However, we were unable to validate the association using a smaller, independent set of patients. Overall, the role of germline variants in metastasis and progression remains unclear. Further analyses with larger datasets including treatment information and functional analysis are needed to better understand the underlying biological processes and the links between this locus and the nearby genes. Prior validation of the reported associations is needed before these findings can be used in clinical-decision making. Therefore, a next step is to study these SNPs in a, preferably, prospective large series of metastasized breast cancer patients. Ultimately, germline variants could help identifying tailored treatments for patients with metastatic disease or better strategies for risk management stratification of aggressive forms of breast cancer.

**Methods**

**Breast cancer samples and genotype data: Breast Cancer Association Consortium (BCAC).** We used genotype and clinico-pathological data (database version 12) from the Breast Cancer Association Consortium (BCAC). The dataset included 1062 breast cancer patients with metastatic primary breast cancer at diagnosis that were genotyped using one of the two different genotyping platforms: iCOGS\textsuperscript{12} and OncoArray\textsuperscript{13}, providing genome-wide coverage of common variants. The main analyses were based on imputed variants using the Haplotype Reference Consortium\textsuperscript{29} as reference panel. All patients were women of European ancestry, aged 26–92 years (median: 60) years with metastasized breast cancer at diagnosis. Women were diagnosed between 1979 and 2014, with a median follow-up was three and a half years. Additional details about the genotype data and sample quality control have been described previously\textsuperscript{7,27,30}. We only analyzed variants that had a minor allele frequency (MAF) >0.01 and an imputation quality \( r^2 > 0.7\) for at least one of the two genotyping platforms (iCOGS or OncoArray). Details about the individual studies included in the analyses, including the array used, associated country and number of patients with metastatic primary breast cancer at diagnosis are given in Supplementary Table 1. The secondary use of data for the study was approved by the Data Access Committee of the BCAC, under the legal provisions of the Memorandum of Understanding and Data Transfer Agreements of Cambridge University which all the contributing institutions, which includes that all contributing institutions provided the data with the appropriate approval of their institutional review boards and informed consent of the participants of the individual studies.

**Statistical and bioinformatic methods.** We estimated the association of the germline variants with breast-cancer specific survival using Cox proportional hazards regression. We analyzed separately the OncoArray and iCOGS datasets and combined the estimates using fixed-effect meta-analyses. Follow-up was right censored on the date of death, last date known alive if death did not occur, or at 15 years after diagnosis, whichever came first\textsuperscript{27}. Time at risk was calculated from the date of diagnosis with left truncation for prevalent cases. The models were stratified by country and included the first two ancestry informative principal components\textsuperscript{12}. We performed the analysis for all breast cancers and for ER-positive and ER-negative tumors separately. To identify evidence of potential cis-regulatory activity, we intersected germline variants with numerous sources of genomic annotation information from primary breast cells (e.g., chromosome conformation, enhancer–promoter correlations, transcription factor and histone modification ChIP-seq). To assess the effect of gene expression on survival we used the Kaplan–Meier plotter on breast tissue data, grade 3 tumors and 15 years of follow-up (180 months)\textsuperscript{18}.

**Validation dataset: SNPs to risk of metastasis (StoRM).** To attempt to validate our results we used data from the SNPs to Risk of Metastasis (StoRM) study. StoRM is a multicentric, prospective, cohort study of metastatic breast cancer patients in France that was originally designed to identify genetic and other factors associated with metastatic relapse and survival\textsuperscript{29}. Patients aged 18 years or older, with a histologically proven breast cancer that was metastatic for less than 1 year were included. All patients that had another coexisting cancer or another cancer diagnosed within the last 5 years, were excluded from the study. Patients were followed from March 2012 to July 2017. Time to progression on the first metastatic treatment was recorded and patients were followed until death, every 6 months for 3 years, and then annually until July 2017. A total of 293 patients were available for the validation. The median follow-up was of 3.2 years. Because of the short total follow-up time (5 years) and the advanced disease stage of the patients in the cohort, both a recorded progression and/or death were considered as an event in the survival analyses. Of the whole set of 293 patients, 239 had a progression and/or died during the follow-up period.

**Ethical approval.** The study was performed in accordance with the Declaration of Helsinki. All individual studies, from which data was used, were approved by the appropriate medical ethical committees and/or institutional review boards. All study participants provided informed consent.

---

[Note: The content is a direct and faithful reproduction of the text as provided, with no additional notes or comments added.]
Author contributions

M.K.S. and M.E.G. conceived the study. M.E.G. performed the main data analyses and drafted the initial manuscript. M.K.S., S.C. and M.E.G. were involved in the interpretation of the data. R.K., Q.W., J.D. and M.K.B. provided database support. J.B. performed the functional analysis. M.K.S. and S.C. worked on revisions of the manuscript. All authors contributed data from their own studies, helped revise the manuscript, and approved the final version. All authors consented to this publication.

Funding

BCAC is funded the European Union’s Horizon 2020 Research and Innovation Programme (Grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and PERSPECTIVE I&I, funded by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l’Économie et de l’Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation. The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report. Additional funding for BCAC is provided via the Confluence project which is funded with intramural funding from the National Cancer Institute Intramural Research Program, National Institutes of Health. Genotyping of the OncoArray was funded by the NIH Grant U19 CA148065, and UK Cancer Grant C1287/A16563 and the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (Grant GPH-129344) and, the Ministère de l’Économie, Science et Innovation du Québec through Genome Québec and the PSRI-IRI-701 grant, and the Quebec Breast Cancer Foundation. Funding for iCOGS 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/A10710, C12292/A1174, 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, and Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. 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 centres in the 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; 2015-7632]. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCC is funded by Cancer Research UK and Breast Cancer Now and acknowledges NHS funding to the NIBR Biomedical Research Centre, and the National Cancer Research Network (NCRN). For the BCFR-PA this work was supported by Grant UM1 CA164920 from the National Cancer Institute. For BIGGS, 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 Centre. The BREOGAN is funded by Acción Estratégica de Salud del Instituto de Salud Carlos III FIS PI12/02125/Cofinanciado FEDER; Acción Estratégica de Salud del Instituto de Salud Carlos III FIS PI17/00918/Cofinanciado FEDER; Acción Estratégica de Salud del Instituto de Salud Carlos III FIS Intrasalud (PI13/01136); Programa Grupos Emergentes, Cancer Genetics Unit, Instituto de Investigación Biomédica Galicia Sur. Xerencia de Xestión Integrada de Vigo-SERGAS, Instituto de Salud Carlos III, Spain; Grant 10CSA012E, Consellería de Industria Programa Sectorial de Investigación Aplicada, PEME I+D e I+D Suma del Plan Gallego de Investigación, Desarrollo e Innovación Tecnológica.
de la Consellería de Industria de la Xunta de Galicia, Spain; Grant EC11-192. Fomento de la Investigación Clínica Independiente, Ministerio de Sanidad, Servicios Sociales e Igualdad, Spain; and Grant FEDER-Innterconecta. Ministerio de Economía y Competitividad, Xunta de Galicia, Spain. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). CCGP is supported by funding from the University of Crete. The CGPS was supported by the Chief Physician Johan Bøserup and Lise Bøserup Fund, the Danish Medical Research Council, and Herlev and Gentofte Hospital. The American Cancer Society funds the creation, maintenance, and updating of the CPS-II cohort. The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by: Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l’Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), Federal Ministry of Education and Research (BMBF) (Germany); the Hellenic Health Foundation, the Stavros Niarchos Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS), PI13/00061 to Granada, PI13/01162 to EPIC-Murcia, Regional Governments of Andalucia, Asturias, Basque Country, Murcia and Navarra, ISCH RETIC (RD06/0020) (Spain); Cancer Research UK (14136 to EPIC-Norfolk; C570/161491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPIC-Oxford) (United Kingdom). The ESTHER study was supported 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 GC-HBCO study is supported by the German Cancer Aid (Grant no 1108158 to M. Vuopio and B. Kühler); The European Regional Development Fund and Free State of Saxony, Germany (LIFE—Leipzig Research Centre for Civilization Diseases, project numbers 713-241202, 14505/2470, 14575/2470). 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 GESBC was supported by the Deutsche Krebshilfe e. V. [70492] and the German Cancer Research Center (DKFZ). The HABCS study was supported by the Claudia von Schilling Foundation for Breast Cancer Research. The HEBCS was financially supported by the Helsinki University Hospital Research Fund, the Finnish Cancer Society, and the Sigrid Juselius Foundation. The HUBCS was supported by the program for supporting the bioresource collections №007-030164/2, and the study was performed as part of the assignment of the Ministry of Science and Higher Education of Russian Federation №2020-220-08-2197. 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 Bert von Kantzows foundation. The KARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. 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. LMBC is supported by the ’Stichting tegen Kanker’. The MARIE study was supported by the Deutsche Krebshilfe e. V. [70-2892-BR L 106332, 108253, 108419, 110826, 110828], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MBCSG is supported by grants from the Italian Association for Cancer Research (AIRC). The MCBCS was supported by the NIH Grants R36CA253187, CA192393, CA116167, CA176758 an NIH Specialized Program of Research Excellence (SPRE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council Grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertainment through theVictorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database. The MEC was supported by NIH Grants CA63464, CA54281, CA098758, CA132839 and CA164973. The MISS study is supported by funding from ERC-2011-294576 Advanced grant, Swedish Cancer Society, Swedish Research Council, Local hospital funds, Berta Kamprad Foundation, Gunnar Nilsson. The MMHS study was supported by NIH Grants CA97396, CA128931, CA116167, CA176758 an NIH Specialized Program of Research Excellence (SPRE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council Grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertainment through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database. The MEC was supported by NIH Grants CA63464, CA54281, CA098758, CA132839 and CA164973. The MISS study is supported by funding from ERC-2011-294576 Advanced grant, Swedish Cancer Society, Swedish Research Council, Local hospital funds, Berta Kamprad Foundation, Gunnar Nilsson. The MMHS study was supported by NIH Grants CA97396, CA128931, CA116167, CA140286 and CA177150. The NBCS has received funding from the K.G. Jebsen Centre for Breast Cancer Research, the Research Council of Norway Grant 193387/590 (to A-L Borresen-Dale and V.N. Kristensen) and Grant 193383/110 (to A-L Borresen-Dale and V.N. Kristensen), South Eastern Norway Health Authority (Grant 39346 to A-L Borresen-Dale) and the Norwegian Cancer Society (to A-L Borresen-Dale and V.N. Kristensen). The NC-BCFR was supported by Grant U11CA164920 from the National Cancer Institute (USA). The CBBCS was supported by the Komen Foundation, the National Cancer Institute (P50 CA058223, U54 CA156733, U01 CA179715), and the North Carolina University Cancer Research Fund. The NHS was supported by NIH Grants P01 CA87969, UM1 CA186107, and U19 CA148065. The NHS2 was supported by NIH Grants UM1 CA176726 and U19 CA148065. 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 ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-RLI). The PBCS was funded by Intramural Research Funds of the National
Cancer Institute, Department of Health and Human Services, USA. Genotyping for PLCO was supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The PLCO is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health. 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. SEARCH is funded by Cancer Research UK [C490/A10124, C490/A16561] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. The University of Cambridge has received salary support for PDPP from the NHS in the East of England through the Clinical Academic Reserve. SKKDKFZS is supported by the DKFZ. The SMC is funded by the Swedish Cancer Foundation and the Swedish Research Council (SIMPLER, VR 2017-00644). STORM was supported by the French Ministry of Health (PHRC K 2011). The SZBCS was supported by Grant PBZ_KBN_122/P05/2004 and the program of the Minister of Science and Higher Education under the name “Regional Initiative of Excellence” in 2019–2022 project number 002/RID/2018/19 amount of financing 12,000,000 PLN. The UCIBCS component of this research was supported by the NIH [CA58860, CA92044] and the Lon V Smith Foundation [LVS39420]. 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.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-99409-3.

Correspondence and requests for materials should be addressed to M.K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021