UNIVERSITY OF
EASTERN FINLAND

GERMLINE COPY NUMBER VARIATION OF *CANCER SUSCEPTIBILITY 8* GENE IN BREAST CANCER RISK AND PROGNOSIS

Henni Elisa Anniina Kosunen
Master of Science thesis
Master´s Degree Programme
in Biomedicine
University of Eastern Finland
Faculty of Health Sciences
School of Medicine
10.6.2019

University of Eastern Finland, Faculty of Health Sciences, School of Medicine
Master´s Degree Programme in Biomedicine
Henni Elisa Anniina Kosunen: Germline Copy Number Variation of *Cancer Susceptibility 8* Gene in Breast Cancer Risk and Prognosis
Master of Science thesis; 36 pages, 0 appendix
Supervisors: Professor PhD Arto Mannermaa & Adjunct Professor PhD Jaana Hartikainen, Institute of Clinical Medicine / Clinical Pathology and Forensic Medicine
10.6.2019

Keywords: breast cancer / *cancer susceptibility 8* / case-control study / germline copy number variation

Abstract

Breast cancer is the most common cancer and cause of cancer mortality in women worldwide, showing substantial variation in its biological and clinical behavior. Germline variants discovered to date can only explain half of the genetic susceptibility to this disease and they are mostly poor prognosticators. Structural copy number variations (CNVs) in the germline DNA are recently identified as missing heritable determinants for breast cancer risk and prognosis through their mediated *cis*-regulation and gene-dosage effects. This is the first case-control study to examine the impact of germline CNV in long non-protein-coding *cancer susceptibility 8* (*CASC8*) gene on breast cancer risk and prognosis among genetically homogenous women (615 invasive cases, 52 *in situ* cases and 146 healthy controls of Northern Savonia/Eastern Finland origin). The number of copies of the *CASC8* gene was determined from the duplex real-time quantitative PCR produced cycle threshold data using single plate and multiplate (combined single plates) analyses. The associations of the copy number status to clinico-pathological (Fisher's exact or Pearson's chi-squared test) and survival data (Kaplan–Meier and covariate-adjusted Cox methods) were assessed. The *CASC8* deletion was significantly associated with breast cancer diagnosis and age of study participants among older women whereas positivity for human epidermal growth factor receptor 2 and involved regional lymph nodes showed the poorest prognosis. A larger cohort of population-representative cases and controls is needed to understand the contribution of this CNV to breast cancer risk. Further investigation of germline CNVs will bring us a step closer to more personalized breast cancer management.

Introduction

Breast cancer is the most frequent cancer and the most common cause of cancer-related death in women worldwide, with an estimated 2.1 million new cases (24.2% of all cancers) and 627,000 deaths (15.0% of all cancer-related deaths) annually based on the GLOBOCAN 2018 statistics (gco.iarc.fr/tomorrow; reviewed by Bray *et al*, 2018). When age-standardized (world) incidence rate per 100,000 women per year in 2018 is compared between more and less economically developed countries, the rate is higher in more developed countries compared to less developed ones (54.4 and 31.3 new cases, respectively), highlighting the differences in the accessibility of early detection, as well as certain lifestyle and reproductive risk factors. However, when the mortality rates are proportioned to the incidence rates, breast cancer survival is more favorable in high-incidence, economically more developed countries (11.6 in contrast to 14.9 deaths). (DeSantis *et al*, 2015; Bray *et al*, 2018) It is extremely disconcerting that the GLOBOCAN 2018 statistics estimate that the worldwide burden of female breast cancer will increase to 3.1 million new cases and 992,000 related deaths per year by 2040 (gco.iarc.fr/tomorrow). This is closely related to the growth and longer life expectancy of the population, the adoption of westernized lifestyles, as well as the changes in reproductive factors in less developed countries (DeSantis *et al*, 2015; Bray *et al*, 2018).

According to the Finnish Cancer Registry from 2017 (cancerregistry.fi/statistics), breast cancer accounts for 30.3% of all invasive cancers among women with around 5,000 new cases annually, meaning that 1 in 8 Finnish women is predicted to develop invasive breast cancer in her lifetime. Based on the same registry, the age-standardized (world, 2017) incidence and mortality rates of breast cancer are 92.5 and 13.0 per 100,000 women annually, respectively. However, standardizing with Finnish population in 2014, the rates are even increased (167.2 and 29.2), reflecting the older age structure and other prominent risk factors known to cause breast cancer. Although mortality has been decreasing as a result of increased awareness and improved screening and treatment strategies (Rampaul *et al*, 2001; DeSantis *et al*, 2015; Ferlay *et al*, 2018), breast cancer is a premature cause of death for more than 900 women annually (15.5% of all cancer-related deaths in 2017; Finnish Cancer Registry). Considering the continuously increased breast cancer incidence in the past decades and its notorious effects on society, there is a vast urge for more effective prevention and treatment strategies (Tao *et al*, 2015).

Breast cancer is a heterogeneous malignancy in terms of the clinical, histological and genetic characteristics of the tumor and thus, its responsiveness to treatments (Elston *et al*, 1999; Rampaul *et al*, 2001). Clinically, the anatomical extent of the disease at the time of diagnosis is described using the TNM classification system where T describes the size of primary tumor, N the number of involved regional lymph nodes and M the presence of distant metastases (Brierley *et al*, 2017). These TNM categories are grouped for tumor staging where stage 0 is proliferated in its original site without invasion of the underlying basement membrane (aka a breast carcinoma *in situ*; ncbi.nlm.nih.gov/mesh/68002278), stage I and II are invaded into surrounding breast tissue, stage III is spread to regional lymph nodes and stage IV is spread to other organs or throughout the body (Brierley *et al*, 2017). Obviously, survival is more favorable in breast cancer that has not invaded the surrounding tissue or notably, sent axillary or distant metastases (Carter *et al*, 1989; Henson *et al*, 1991; Clayton & Hopkins, 1993). With respect to this invasion capability of the cancer cells and prognosis of the breast cancer patient, tumor grade is defined by the degree of mammary epithelial differentiation, nuclear pleio-morphism and mitotic counts (Elston & Ellis, 1991). Many studies have proved that survival time is longer in breast cancer patients with well-differentiated lower-grade malignancy compared to patients with poorly-differentiated higher-grade malignancy (Davis *et al*, 1986; Elston & Ellis, 1991; Henson *et al*, 1991; Clayton & Hopkins, 1993; Rakha *et al*, 2008).

The epithelial tumors of the breast can be classified into *in situ* and invasive carcinomas based on their histological architecture. Approximately 75% of invasive breast carcinomas are more aggressive ductal carcinomas not otherwise specified whereas the rest 25% are histologically special types, such as tubular, lobular, mucinous, medullary, as well as invasive cribriform and papillary carcinomas with more favorable prognoses. (Elston *et al*, 1999; Dawson *et al*, 2013) The clinical management of patients with breast cancer involves testing of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) status by immunohistochemistry or fluorescence *in situ* hybridization (*erb-b2 receptor tyrosine kinase 2*, the encoding gene for HER2) (Allred, 2010; Dawson *et al*, 2013). Nuclear transcription factors ER and PR mediate their steroid ligand-stimulated proliferation of breast epithelial cells and are expressed in around 75% and 65% of invasive breast carcinomas, respectively. These tumors are responsive to hormonal therapies, including tamoxifen and aromatase inhibitors. In contrast, HER2 is located on the surface of breast epithelial cells where it promotes cell proliferation and survival. Its over-expression and/or amplification is observed in around 15% of invasive breast carcinomas which are treated with antibody-based therapies, such as trastuzumab. (Allred, 2010) Considering the gene expression patterns, luminal A and luminal B subtypes are ER-positive breast cancers which are separated based on increased HER2 and/or cell proliferation marker (e.g. cyclin B1 and Ki-67) expression in more aggressive luminal B subtype (Cheang *et al*, 2009; Caldarella *et al*, 2011). ER-negative breast cancers are either positive (HER2 over-expressing) or negative (triple-negative basal-like) for HER2 or they express genes typical of adipose/non-epithelial cells (normal breast tissue-like) (Perou *et al*, 2000; Sørlie *et al*, 2001). HER2 over-expressing and triple-negative basal-like subtypes have been shown to associate with poorer prognosis than the luminal and normal breast tissue-like subtypes (Sørlie *et al*, 2001; Caldarella *et al*, 2011). In summary, both clinical, histological and genetic characteristics of the tumor are important predictive and prognostic factors in managing patients with breast cancer (Elston *et al*, 1999; Rampaul *et al*, 2001; Tao *et al*, 2015)

High breast cancer incidence in Finland as in other economically more developed countries and continuously rising incidence in less developed ones arise from the ageing of the population, as well as the western diet, sedentary lifestyles and changed reproductive factors associated with family planning (DeSantis *et al*, 2015; Bray *et al*, 2018). The main risk factors for the development of breast cancer involve female sex, middle or older age and first-degree family history of breast cancer. Numerous reproductive risk factors, such as early menarche, late first full-term pregnancy or nulliparity, short lactation, late menopause and use of oral contraceptives or postmenopausal hormone replacement therapies, elevate the risk of breast cancer due to an increased exposure to estrogen. Obesity also increases estrogen levels via the peripheral conversion of androgens and the low level of sex hormone-binding globulin. Tumorigenesis in breast epithelial cells is further increased by adipose tissue expressed pro-inflammatory and pro-angiogenic cytokines. (Rojas & Stuckey, 2016) Therefore, body fatness (Cui *et al*, 2002; Guo *et al*, 2017) is a predictive and prognostic factor whereas physical activity (Steindorf *et al*, 2013) is a protective factor for breast cancer. Both active and passive smoking, as well as alcohol consumption may slightly increase the risk of breast cancer, perhaps due to carcinogens and carcinogenic metabolites which damage DNA and disrupt estrogen metabolism. The adverse effect is suggested to be more pronounced in less-differentiated breast tissue of young woman before her first full-term pregnancy. However, more research is required to confirm the associations of smoking and alcohol to breast cancer risk. (Dossus *et al*, 2014; Romieu *et al*, 2015)

In contrast to relatively common lifestyle and reproductive risk factors, inherited susceptibility is considered to account less than 10% of all breast cancer cases. The most common germline high-penetrance mutations are in *DNA repair associated 1* (*BRCA1*) and *2* (*BRCA2*) genes which increase the relative risk of developing breast cancer around 10-fold. Compared to non-familial cases, most of familial breast cancers caused by mutations in these genes tend to be poorly-differentiated higher-grade triple-negative basal-like (*BRCA1*) and luminal (*BRCA2*) subtypes diagnosed before menopause. Mutations in the tumor suppressor genes *tumor protein p53* and *phosphatase and tensin homolog* in rare familial Li Fraumeni and Cowden Syndrome cancer diseases confer even higher risks for breast cancer development. (Palacios *et al*, 2008; Rojas & Stuckey, 2016) However, only few of breast cancer families exhibit these germline gene mutations, so many variants with a smaller effect play a role in familial, as well as in non-familial breast cancer risk (Ghoussaini *et al*, 2013). Intermediate-penetrance genes *checkpoint kinase 2, ATM serine/threonine kinase, BRCA1 interacting protein C-terminal helicase 1, partner and localizer of BRCA2* and *RAD50 double strand break repair protein* promote DNA repair by regulating or interacting with the BRCA protein complex. Rare mutations in these genes have been reported to increase breast cancer risk around 2-fold (CHEK2 Breast Cancer Case-Control Consortium, 2004; Renwick *et al*, 2006; Seal *et al*, 2006; Rahman *et al*, 2007), *RAD50 double strand break repair protein* and *partner and localizer of BRCA2* gene mutations conferring a 4-fold increased risk among genetically isolated homogeneous Finns (Heikkinen *et al*, 2006; Erkko *et al*, 2007). Apart from previous genome-wide linkage and mutational screening studies, genome-wide association studies (GWASs) have discovered about hundred common variants associated with low-penetrance breast cancer susceptibility with a < 1.5-fold relative risk. These genes are involved in pathways relevant to tumorigenesis, including cell division, proliferation and apoptosis, as well as DNA repair. (Ghoussaini *et al*, 2013)

Although germline mutations and variants associated with an increased breast cancer risk have been widely analyzed (Ghoussaini *et al*, 2013), the impact of germline copy number variations (CNVs) on breast cancer susceptibility and prognosis is poorly understood. Germline CNVs, referred as the parental-originated DNA segments of 50 bp to 1 Mb in size copied at different efficiency between individuals in meiosis, comprise around 10% of the human genome. (Kumaran *et al*, 2017) More than 552,000 CNVs among healthy individuals of diverse ethnicities have been reported in the Database of Genomic Variants since 2016 (dgv.tcag.ca/ dgv/app/home; Zarrei *et al*, 2015). Apart from increasing the genetic diversity between individuals, germline CNVs have been reported to have predictive and prognostic values to a wide-range of epithelial cancers affecting colon, ovaries and prostate since duplication, deletion or inversion event can disrupt the coding sequence or regulatory region of the gene involved in cell differentiation, proliferation or survival. Moreover, the CNV event can change the expression level of the target gene in a tissue-specific manner and thus contribute to cancer development. By activating proto-oncogenes, inactivating tumor suppressor genes or altering other ways the pathways relevant to tumorigenesis, germline CNVs contribute the missing susceptibility of breast cancer that could not been discovered by the GWASs. However, only few of these CNVs have been found so far, in which fewer with prognostic relevance. (Kumaran *et al*, 2017)

As examples of common germline CNVs contributing a low breast cancer susceptibility with a minor allele frequency > 5% in population, the copy number (CN) deletion in exon 4 of *microtubule associated scaffold protein 1* has been shown to enhance mitochondrial tumor suppressor function of this gene and to inhibit epidermal growth factor signaling, thus be associated with a decreased familial breast cancer risk among German women without high-penetrance mutations in *BRCA1* and *BRCA2* genes (Frank *et al*, 2007). The fusion gene of *apolipoprotein B mRNA editing enzyme catalytic subunits 3A* and *3B* due to CN deletion within intervening coding region has been shown to relate to an increased breast cancer risk among European and Chinese women. Predisposition mechanism is thought to be an increased DNA mutation by deamination of (5-methyl)-cytosine to difficulty repaired thymine than having interactions with other environmental or genetic risk factors. This CNV is almost 3-fold more common among Chinese than European women. (Xuan *et al*, 2013) Upon exposure to estrogen, the CN deletion in enhancer region on the chromosome 2q35 has been investigated to decrease the expression of the downstream *insulin like growth factor binding protein 5* gene via a large chromatin loop. This enhancer CNV is thus thought to relate to a decreased insulin-like growth factor signaling and breast cancer risk among European American and African American women. (Wyszynski *et al*, 2016) The CN deletions in detoxification and metabolizing enzymes *glutathione S-transferase theta 1* and *UDP glucuronosyltransferase family 2 member B17* genes have been shown to cover an increased breast cancer risk by changing gene expression in breast tissue by a dosage-effect. The same study also discovered hundreds of candidate CNVs associated with breast cancer risk, some of which had prognostic significance, but they must be validated in non-Caucasians in order to confirm the findings. (Kumaran *et al*, 2017)

In contrast to common CNVs, rare germline CNVs with a minor allele frequency < 1% in population have been reported to confer high susceptibilities for familial and early-onset breast cancer development in patients negative for known high-penetrance gene mutations. Pylkäs and colleagues (2012) have found that the genes harbored with rare CNVs were associated with the maintenance of genomic integrity centered on tumor suppressor p53 function, as well as the signaling and metabolism of biologically active estrogen, β-estradiol. For example, the CNVs were found in genes *BLM RecQ like helicase*, *RecQ like helicase 4* and *DNA cross-link repair 1C* which respond and repair DNA damage, *estrogen receptor 2* that encodes an ER for β-estradiol and *cytochrome P450 family 2 subfamily C member 19* that encodes a metabolic enzyme for β-estradiol. (Pylkäs *et al*, 2012) Masson and colleagues (2014) have found rare CNVs in previously cancer-associated genes *caspase recruitment domain family member 11* that interacts with pro-apoptotic proteins, *MLLT11 transcription factor 7 cofactor* that is involved in breast cancer metastasis, *protein tyrosine kinase 2 β* that regulates cell's actin-cytoskeleton rearrangement and cell adhesion, migration and spreading, *Rho GTPase activating protein 26* and *Rho guanine nucleotide exchange factor 12* which regulate Rho-dependent cell's actin-cytoskeleton organization and G protein-coupled receptor signaling, as well as *fragile histidine triad* and *WW domain containing oxidoreductase* which are tumor suppressor genes in cancer-associated fragile sites. Moreover, the genes *replication protein A3*, *nibrin* and *MRE11 homolog double strand break repair nuclease* harbored with rare CNVs have been shown to maintain the genomic integrity. (Masson *et al*, 2014; functions provided by NCBI Entrez Gene [ncbi.nlm.nih.gov/gene]) By disrupting important biological networks which control normal cell functions, the rare germline CNVs are thought to change the behavior of breast epithelial cells toward tumorigenesis. (Pylkäs *et al*, 2012; Masson *et al*, 2014)

Because common and rare germline CNVs cover a higher portion of the genome than other inherited variants described as single-nucleotide polymorphisms (SNPs) discovered by the GWASs (Ghoussaini *et al*, 2013; Zarrei *et al*, 2015; Kumaran *et al*, 2017), a detailed knowledge of the germline CNVs would not only provide further insights into the development of breast cancer but would also help us to develop more personalized strategies for the early diagnosis and treatment of this malignancy (Kuiper *et al*, 2010). The germline CNVs are captured to date using a wide-range of microarray- and sequencing-based approaches with different resolutions, such as comparative genomic hybridization, SNP and oligonucleotide arrays, as well as Sanger and next-generation sequencing techniques (Zarrei *et al*, 2015).

The gene desert region of 1.18 Mb in size on chromosome 8q24.21 has been shown to contain multiple independent risk variants conferring susceptibilities for breast, colon, ovarian and prostate cancers, some of which are commonly associated with all these cancers except of breast cancer (Ghoussaini *et al,* 2008). Considering only breast cancer risk among European women, five independently risk-associated variants within this gene desert region have been found to date, all of which are associated with the luminal subtype of breast cancer. These variants have been thought to harbor the long-range regulatory elements of well-known *proto-oncogene c-MYC* that regulates the transcription of several genes involved in cell proliferation, growth and survival and thus, is frequently amplified and/or over-expressed in breast tumors. (Hynes & Stoelzle, 2009; Shi *et al*, 2016) One candidate long non-protein-coding RNA (lncRNA) gene within this *c-MYC* regulatory region is a *cancer susceptibility 8* (*CASC8*) because of it contains 2 out of 5 these breast cancer risk-associated variants and moreover other common variants known to be associated with colorectal, gastric and prostate cancers (Cui *et al*, 2018). Considering that germline high-penetrance mutations and low-to-intermediate risk-associated variants discovered to date can only explain half of the genetic predisposition to breast cancer (Ghoussaini *et al*, 2013; Kumaran *et al*, 2017), and specific genetic risk factors exist among relatively stable and genetically isolated populations like us Finns (Hartikainen *et al*, 2005), the germline CNV in *CASC8* may play a role in breast cancer susceptibility and/or prognosis among women of Northern Savonia/Eastern Finland origin.

This study involved a total of 615 cases with invasive breast carcinoma, 52 cases with *in situ* breast carcinoma and 146 healthy controls, all analyzed by two different CN methods. The CN deletion located in the *CASC8* gene was shown to be significantly associated with breast cancer diagnosis among older women and it was more often found in older invasive breast cancer cases and healthy controls. There was no evidence of an association between the *CASC8* CN deletion and the clinico-pathological characteristics (histology, size, stage, grade, lymph node status, distant metastasis, local/distant relapse and hormone receptor status [ER, PR, HER2]) of breast tumor. HER2-positive disease and regional lymph node involvement were shown to be more significant prognostic factors for breast cancer survival than investigated *CASC8* structural variant. These findings are consistent with current knowledge that structural CNVs consisting of gains and losses of considerable length genomic stretches in the germline DNA contribute to breast cancer susceptibility and their role in its pathogenesis requires further investigations.

Materials and methods

Study population

The study contained 813 participants from previous Kuopio Breast Cancer Project (KBCP) and Itä–Länsi Breast Cancer Project (ILRS), including 615 cases with invasive breast carcinoma, 52 cases with *in situ* breast carcinoma and 146 healthy controls, all came from the Northern Savonia of Eastern Finland. In the KBCP, the cases were diagnosed between April 1990 and December 1995 in the Kuopio University Hospital and their long-term area of residence- (urban/rural) and age- (within ± 5 years) individually matched controls were obtained from the National Population Register (Hartikainen *et al*, 2005). The ILRS only included cases diagnosed with breast cancer between May 2011 and December 2014 in the Kuopio University Hospital (unpublished project material collected by Prof. Arto Mannermaa's research group). Leukocyte genomic DNA (gDNA) of blood samples from all participants had been extracted using standard chloroform-phenol extraction method (KBCP) or commercially available DNA extraction kit (Qiagen QIAamp DNA Blood Midi Kit, Hilden, Germany; ILRS). Clinical and follow-up (survival) data from the cases had been collected from hospital records. Both KBCP and ILRS have been advocated by the joint Research Ethics Committee of the University of Eastern Finland and the Kuopio University Hospital when written informed consents have also been given by all study participants.

Sample preparation

The quality (A_{260}/A_{280}) and concentration of gDNA were measured from all samples which had not been measured recently using the NanoDrop ND-1000 UV/Vis Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). After that, ~ 5 ng/µL dilutions were made using PCR-grade water. If the first real-time quantitative PCR amplification did not succeed (the A_{260}/A_{280} ratio and the concentration of gDNA were not within the recommended ranges), the Invitrogen^{TM} Qubit® 3.0 Fluorometer and dsDNA BR Assay Kit (Life Technologies Corporation, Eugene, OR, USA) were used for the re-quantitation of double-stranded gDNA in samples followed by their more accurate dilution and new PCR amplification.

Real-time quantitative PCR amplification

The germline CNV in *CASC8* was examined by a duplex real-time quantitative PCR using the pre-designed Applied Biosystems TaqMan® Copy Number Assays (Life Technologies

Corporation, Pleasanton, CA, USA). Primers and FAM^{TM} dye-labeled MGB probes were selected to target *CASC8* (assay ID Hs06219825_cn, build on GRCh38) based on the previous findings of the research group (Prof. Arto Mannermaa, personal communications; build on GRCh37 and converted using the Galaxy Lift-Over version 1.0.6 tool [usegalaxy.org]). The VIC® dye-labeled TAMRA^{TM} -probed TaqMan® Copy Number Reference Assay RNase P was used as an endogenous control since its target gene *ribonuclease P RNA component H1* (*RPPH1*) exists in one copy on each chromosome 14q11.2. The Applied Biosystems TaqMan® Universal PCR Master Mix without AmpErase® UNG (Life Technologies LTD, Woolston, WA, UK) provided AmpliTaq Gold® DNA polymerases and dNTPs for the real-time PCR reactions. Each reaction consisted of 10 µL 2X TaqMan® Universal PCR Master Mix, 1 µL both 20X TaqMan® Copy Number Assays, 4 µL (~ 20 ng) gDNA samples and 4 µL PCR-grade water. A gDNA sample with previously identified CN deletion in *CASC8* by the research group was used as a calibrator for data normalization. All samples were amplified in triplicates instead of recommended quadruplets using the LightCycler® 96 Instrument (Roche Diagnostics GmbH, Mannheim, Germany) and the universal cycling conditions (10 min at 95°C for 1 cycle followed by 40 cycles of 15 sec at 95°C and 1 min at 60°C).

Data analysis

The germline CNV in *CASC8* was examined from cycle threshold (C_T) raw data produced by the LightCycler® 96 Application version 1.1.0.1320 software (Roche Diagnostics GmbH, Mannheim, Germany). The number of *CASC8* copies in each gDNA sample replicate group was calculated using two different CN methods since it was not known which would be better: 1) A single plate analysis using the Microsoft Excel software (Microsoft Corporation, Redmond, WA, USA); 2) A multiplate analysis where data from multiple plates sharing the same PCR setup were combined for analysis using the Applied Biosystems CopyCaller version 2.0 software (Life Technologies Corporation, Carlsbad, CA, USA). Both CN analyses used the comparative C_T method ($\Delta\Delta C_T$) of relative quantitation:

$$\Delta\Delta C_T = \mu(C_{T(\mathit{CASC8})} - C_{T(\mathit{RPPH1})}) \text{ gDNA replicates} - \mu(C_{T(\mathit{CASC8})} - C_{T(\mathit{RPPH1})}) \text{ calibrator replicates}$$

➔ $CN_{\mathit{CASC8}} = CN(1)_{\text{calibrator}} \times 2^{(-\Delta\Delta CT)}$ in each gDNA sample replicate group.

Outliers were excluded based on the default settings for the Applied Biosystems TaqMan® Copy Number Assays (VIC $C_T > 32$, FAM $C_T > 40$ and/or $\Delta C_T > 4.0$) and other settings which differed between the CN analysis methods: 1) One outlier C_T value was excluded from the single plate analysis if the standard deviation (SD) between gDNA sample triplicates was > 0.3; 2) Any sample with the absolute z-score value > 2.65 and/or the confidence value < 0.9 was excluded from the multiplate analysis since the CN call was not reliable. The CN values for *CASC8* were rounded as follows: 1) In the single plate analysis using more accurate ranges: 0.65–1.34 (CN 1 = deletion), 1.65–2.34 (CN 2 = diploid) and 2.65–3.34 (CN 3 = duplication); 2) In the multiplate analysis using less accurate ranges: 0.50–1.49 (CN 1), 1.50–2.49 (CN 2) and 2.50–3.49 (CN 3). The gDNA samples which did not meet these quality demands were considered as having unknown (X) CN and were excluded from statistical analyses.

Statistical analysis

Statistical analyses were carried out using the SPSS® Statistics version 25.0 software (IBM, Armonk, NY, USA). Study data was normally distributed and thus appropriate for parametric testing based on the Kolmogorov–Smirnov test. The Fisher's exact test was used to estimate the differences in CN frequencies between breast cancer cases and healthy controls. The Pearson's chi-squared test was used to estimate the association between different CNs and available clinical parameters (age at breast cancer diagnosis or interview of control, as well as the histology, size, stage, grade, lymph node status, distant metastasis, local/distant relapse and hormone receptor status [ER, PR, HER2, triple negativity] of invasive breast tumor). However, the Fisher's exact test was used if some of the data cells contained < 5 expected counts. The distribution of ages among the CN status was represented by the histograms in order to confirm the results from age-CN associations. The univariate Kaplan–Meier method was used to estimate the effect of different CNs on breast cancer-specific survival (BCSS) and relapse-free survival (RFS) times and the difference in survival among the CN status was tested using the log-rank Mantel–Cox test. Hazard ratios (HRs) were assessed using the multivariate Cox proportional hazards model by adjusting for major clinico-pathological covariates. In both survival analyses, cases with existed metastases at the time of breast cancer diagnosis were excluded (BCSS & RFS) and cases with breast cancer-unrelated death/breast cancer survival (BCSS) or getting no relapse/new breast cancer diagnosis (RFS) were censored. The survival rates were represented with their 95% confidence intervals (CIs). All statistical analyses were 2-sided with a 5% type I error rate and *P*-values ≤ 0.05 were considered as statistically significant.

Results

The frequency of CASC8 CN deletion did not differ between breast cancer cases and healthy controls

The real-time quantitative PCR consisted of gDNA samples from 615 cases with invasive breast carcinoma, from 52 cases with *in situ* breast carcinoma and from 146 healthy controls. Due to high SD between ΔC_T values of gDNA sample replicates or improper quality metrics of analyzed gDNA sample, the CN status of *CASC8* could not be determined for 91 participants (37 invasive cases, 3 *in situ* cases and 51 controls; 11.2%) using the single plate analysis (Fig. 1A) and for 10 participants (2 invasive cases and 8 controls; 1.2%) when using the multiplate analysis (Fig. 1B). In the single plate analysis, 28 participants (24 invasive cases, 2 *in situ* cases and 2 controls) had one copy of *CASC8* (3.9%) and 694 participants (554 invasive cases, 47 *in situ* cases and 93 controls) had two copies of *CASC8* (96.1%). In contrast, the multiplate analysis revealed 27 participants (23 invasive cases, 2 *in situ* cases and 2 controls) having one gene copy (3.4%) and 775 participants (589 invasive cases, 50 *in situ* cases and 136 controls) having diploid gene copy (96.6%). The multiplate analysis also revealed 1 invasive case that had three copies of *CASC8* (0.1%), but this was excluded from further statistical analyses due to its biased effect on the results. Overall, the CN frequencies did not differ between invasive breast cancer cases and healthy controls using either method (the Fisher's exact test *P*-values 0.563 and 0.204 for the single plate and the multiplate analyses, respectively). The inclusion of *in situ* cases did not affect the results (respective *P*-values 0.566 and 0.204).

The CN deletion in CASC8 was associated with age at breast cancer diagnosis

The CN deletion in *CASC8* was associated with the age of invasive cases at the time of breast cancer diagnosis (Table 1). Using the 10-year age category (≤ 39, 40–49, 50–59, 60–69 and 70 ≥), *P*-values were 0.010 for the single plate and 0.020 for the multiplate analyses evaluated by the Fisher's exact test. CN status was also significant for cases below or equal/over 60 years of age using both analysis methods (the Pearson's chi-squared $P = 0.018$ and 0.029). In contrast to invasive cases, the CN deletion was not associated with the age of controls at interview with corresponding Fisher's exact *P*-values 0.184 and 0.435 for the single plate, and 0.237 and 0.394 for the multiplate analyses. Also, the clinico-pathological characteristics of tumor at the time of diagnosis (histology, size, stage of invasion, grade of differentiation, number of involved regional lymph nodes, having distant metastases or local/distant relapse, as well as hormone receptor status) were not statistically significant for CN status using either analysis method. The lowest non-significant *P*-values were observed for well-differentiated lower-grade tumors (Pearson's chi-squared *P* = 0.145 and 0.131). However, clinical data was not available from all participants as also summarized in Table 1.

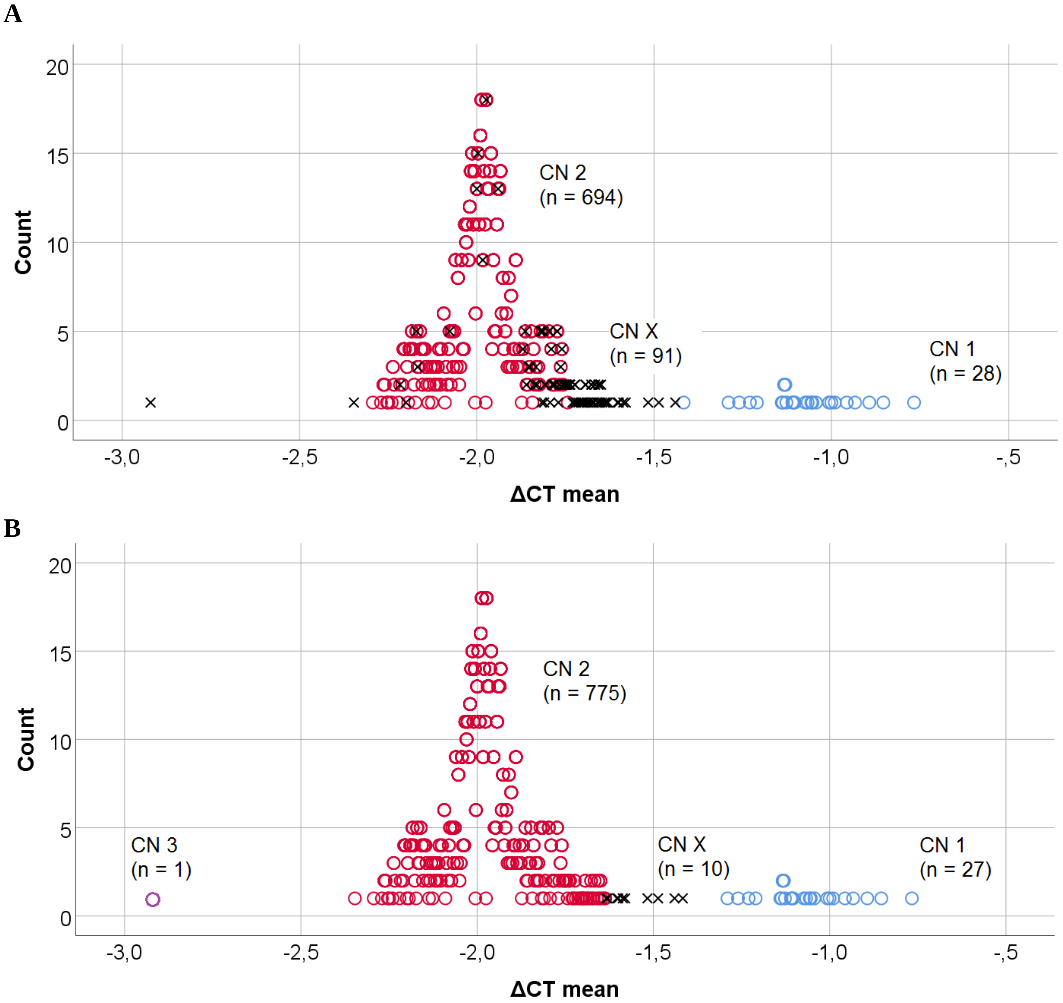


Figure 1. Distributions of mean ΔC_T values for a specific CN of *CASC8*. The count of gDNA samples having a particular ΔC_T value is displayed in Y axis and the mean ΔC_T value for the technical replicates of the associated gDNA sample is shown in X axis. The different CNs are marked as follows: blue (deletion/one copy of *CASC8*), red (diploid/two copies of *CASC8*), purple (duplication/three copies of *CASC8*) and black (unknown copies of *CASC8*). (**A**) ΔC_T plot from single plate analysis. The gDNA samples having unknown CNs for *CASC8* were excluded from further statistical analyses since their technical replicates exhibited broad range of ΔC_T values. (**B**) ΔC_T plot from multiplate analysis. The gDNA samples which had either insufficient quality for reliable CN assessment based on the absolute z-score and/or the confidence values or three copies of *CASC8* were excluded from further statistical analyses.

Table 1. Association of the germline CNV in *CASC8* and clinical parameters

Clinical parameters	Single plate analysis				Multiplate analysis			
	N[a]	CN 1 N (%)	CN 2 N (%)	*P*- value	N[a]	CN 1 N (%)	CN 2 N (%)	*P*- value
Age of controls								
≤39, 40-49, 50-59, 60-69, 70≥	93/146	2 (2.2)	91 (97.8)	0.184[b]	136/146	2 (1.5)	134 (98.5)	0.237[b]
<60 vs. 60≥	93/146	2 (2.2)	91 (97.8)	0.435[b]	136/146	2 (1.5)	134 (98.5)	0.394[b]
Age of invasive cases								
≤39, 40-49, 50-59, 60-69, 70≥	394/615	19 (4.8)	375 (95.2)	**0.010[b]**	428/615	18 (4.2)	410 (95.8)	**0.020[b]**
<60 vs. 60≥	394/615	19 (4.8)	375 (95.2)	**0.018**	428/615	18 (4.2)	410 (95.8)	**0.029**
Tumor histology								
ductal vs. lobular vs. others	574/615	24 (4.2)	550 (95.8)	0.734[b]	608/615	23 (3.8)	585 (96.2)	0.648[b]
ductal vs. others	574/615	24 (4.2)	550 (95.8)	0.654	608/615	23 (3.8)	585 (96.2)	0.493
invasive ca vs. *in situ* ca	623/667	26 (4.2)	597 (95.8)	1.000[b]	660/667	25 (3.8)	635 (96.2)	1.000[b]
Tumor size								
T1 vs. T2 vs. T3 vs. T4	574/615	24 (4.2)	550 (95.8)	0.689[b]	608/615	23 (3.8)	585 (96.2)	0.790[b]
T4 vs. others	574/615	24 (4.2)	550 (95.8)	0.404[b]	608/615	23 (3.8)	585 (96.2)	0.465[b]
Tumor stage								
I vs. II vs. III vs. IV	387/615	18 (4.7)	369 (95.3)	0.828[b]	420/615	17 (4.0)	403 (96.0)	0.875[b]
I vs. others	387/615	18 (4.7)	369 (95.3)	0.811	420/615	17 (4.0)	403 (96.0)	1.000
Tumor grade								
1 vs. 2 vs. 3+4	570/615	23 (4.0)	547 (96.0)	0.281	604/615	22 (3.6)	582 (96.4)	0.233
1 vs. others	570/615	23 (4.0)	547 (96.0)	0.145	604/615	22 (3.6)	582 (96.4)	0.131
Lymph node status								
N0 vs. N1 vs. N2 vs. N3	566/615	23 (4.1)	543 (95.9)	0.403[b]	599/615	22 (3.7)	577 (96.3)	0.416[b]
N0 vs. others	566/615	23 (4.1)	543 (95.9)	0.388	599/615	22 (3.7)	577 (96.3)	0.387
Distant metastasis								
yes vs. no	393/615	19 (4.8)	374 (95.2)	0.802	426/615	18 (4.2)	408 (95.8)	0.802
Local/distant relapse								
yes vs. no	575/615	24 (4.2)	551 (95.8)	0.655	609/615	23 (3.8)	586 (96.2)	0.369
ER status								
positive vs. negative	557/615	23 (4.1)	534 (95.9)	0.578[b]	591/615	22 (3.7)	569 (96.3)	0.578[b]
PR status								
positive vs. negative	555/615	23 (4.1)	532 (95.9)	0.822	589/615	22 (3.7)	567 (96.3)	1.000
HER2 status								
positive vs. negative	536/615	23 (4.3)	513 (95.7)	1.000[b]	566/615	22 (3.9)	544 (96.1)	1.000[b]
ER, PR and HER2 negativity								
yes vs. no	552/615	23 (4.2)	529 (95.8)	0.461[b]	586/615	22 (3.8)	564 (96.2)	0.457[b]

a: N = 615 (all cases with invasive carcinoma), N = 667 (all cases with invasive or *in situ* carcinoma) and N = 146 (all healthy controls)

b: Cells had expected count < 5, so the Fisher's exact test was applied. Otherwise *P*-values were calculated using the Pearson's chi-squared test.

The CN deletion in CASC8 was associated with age of study participants

To evaluate further the association between CN status and age of study participants, the Pearson's chi-squared and the Fisher's exact tests were carried out for all available invasive cases and healthy controls (Table 2). The CN deletion in *CASC8* was observed to associate with age of invasive cases and healthy controls: *P*-values were 0.010 and 0.015 for the single plate and the multiplate analyses using the 10-year age category, but when using the age below or equal/over 60 years, the *P*-value from multiplate analysis was even more significant (*P* = 0.009). As in the case of clinico-pathological characteristics of breast tumor (Table 1), information about age was missing from many study participants, also summarized in Table 2.

Table 2. Association of the germline CNV in *CASC8* and age of study participants

Age of invasive cases and healthy controls	Single plate analysis				Multiplate analysis			
	N[a]	CN 1 N (%)	CN 2 N (%)	*P*-value	N[a]	CN 1 N (%)	CN 2 N (%)	*P*-value
≤39, 40-49, 50-59, 60-69, 70≥	487/761	21 (4.3)	466 (95.7)	**0.010**[b]	564/761	20 (3.5)	544 (96.5)	**0.015**[b]
<60 vs. 60≥				**0.010**				**0.009**

a: N = 761 (all cases with invasive carcinoma, as well as healthy controls)
b: Cells had expected count < 5, so the Fisher's exact test was applied. Otherwise *P*-values were calculated using the Pearson's chi-squared test.

The CN deletion in CASC8 was associated with breast cancer diagnosis among older women and it was more often found in older study participants

The distribution of ages among the different CN status of *CASC8* was next evaluated by histograms in order to confirm the previous findings of the germline CNV association with age at breast cancer diagnosis (Table 1) and age of invasive cases and healthy controls (Table 2). Invasive cases having copy deletion were diagnosed with breast cancer at an older age than diploid CN cases: The mean and median of ages were 63.00 (95% CI = 57.43–68.57)/57.90 (56.44–59.35) and 64.00/55.00 for cases involved in the single plate analysis (Fig. 2A) while 62.67 (56.80–68.53)/57.79 (56.39–59.18) and 63.50/55.00 for cases in the multiplate analysis (Fig. 2B). Age SDs were also lower among invasive cases with copy deletion versus diploid CN: SD = 11.55 vs. 14.34 using the single plate analysis and SD = 11.79 vs. 14.39 using the multiplate analysis. When healthy controls were included in the histograms (Fig. 3A–B), the mean and median of ages were slightly reduced in both CN groups, this trend being similar to different CN analysis methods. Again, the study participants having one gene copy of

CASC8 were older than their diploid CN counterparts and their age SDs were also lower. In conclusion, the CN deletion in *CASC8* was associated with breast cancer diagnosis among older women and it was more often found in older invasive cases and healthy controls.

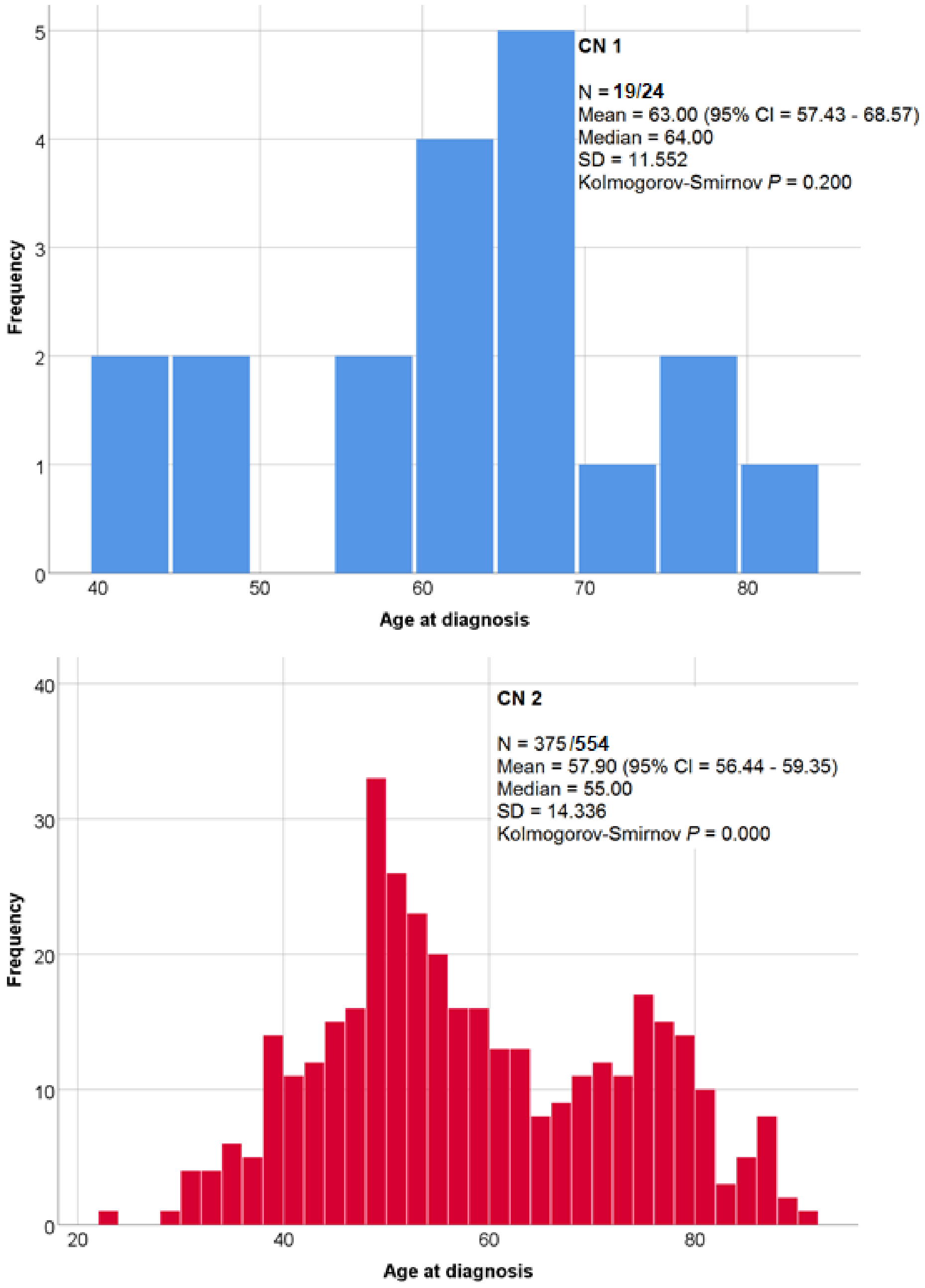


Figure 2A. Age distribution of invasive breast cancer cases among the single plate analyzed CN status. The frequency of invasive cases is displayed in Y axis and the age at the time of diagnosis is shown in X axis. Descriptive values are also depicted. N = 24 (all invasive cases having one copy of *CASC8* [blue]); N = 554 (all invasive cases having two copies of *CASC8* [red]). *P*-values for association of the germline CNV in *CASC8* and age at the time of diagnosis are 0.010 (ages of ≤ 39, 40–49, 50–59, 60–69 and 70 ≥, evaluated by the Fisher’s exact test) and 0.018 (ages of < 60 and 60 ≥, evaluated by the Pearson’s chi-squared test).

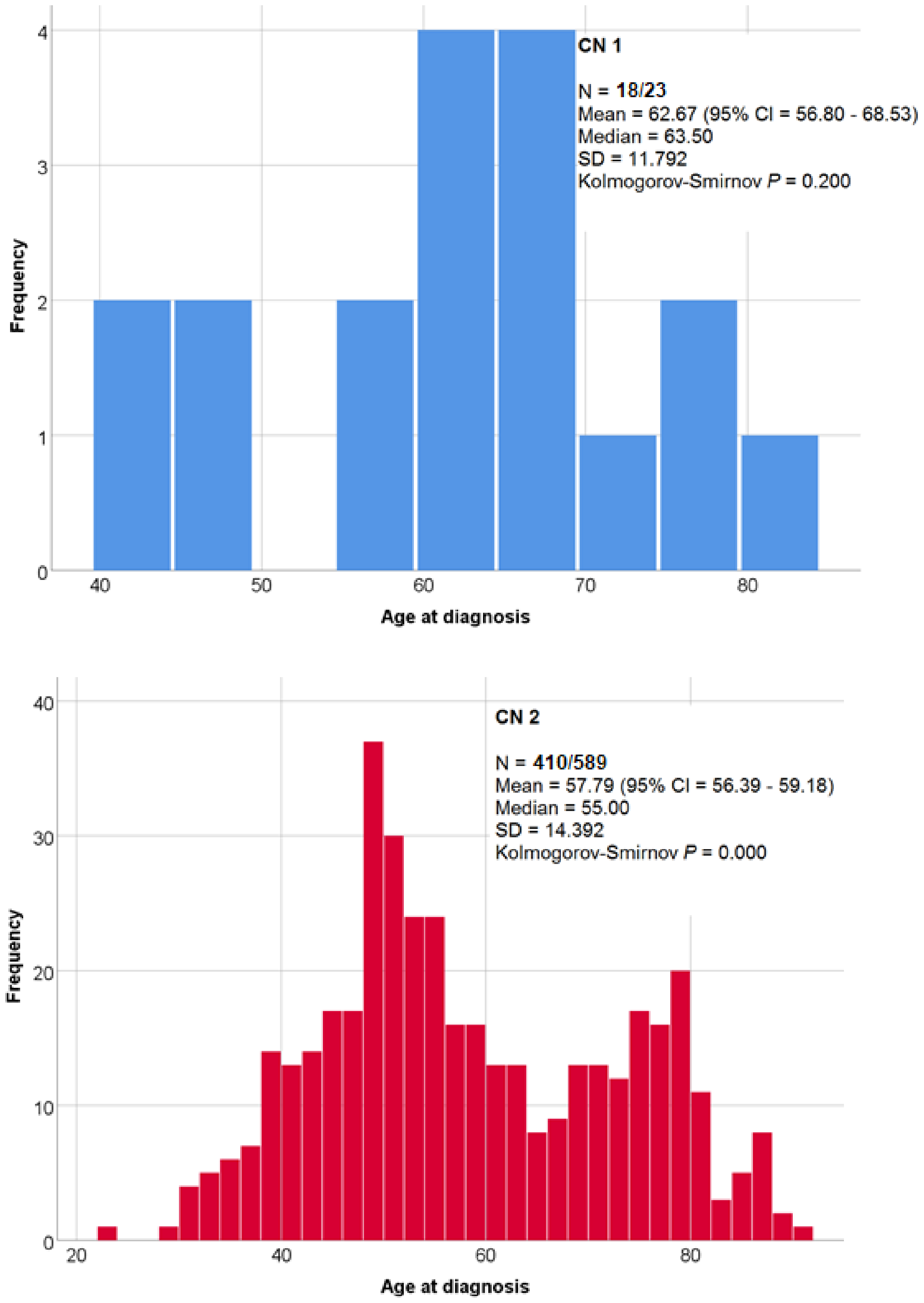


Figure 2B. Age distribution of invasive breast cancer cases among the multiplate analyzed CN status. The frequency of invasive cases is displayed in Y axis and the age at the time of diagnosis is shown in X axis. Descriptive values are also depicted. N = 23 (all invasive cases having one copy of *CASC8* [blue]); N = 589 (all invasive cases having two copies of *CASC8* [red]). *P*-values for association of the germline CNV in *CASC8* and age at the time of diagnosis are 0.020 (ages of ≤ 39, 40–49, 50–59, 60–69 and 70 ≥, evaluated by the Fisher's exact test) and 0.029 (ages of < 60 and 60 ≥, evaluated by the Pearson's chi-squared test).

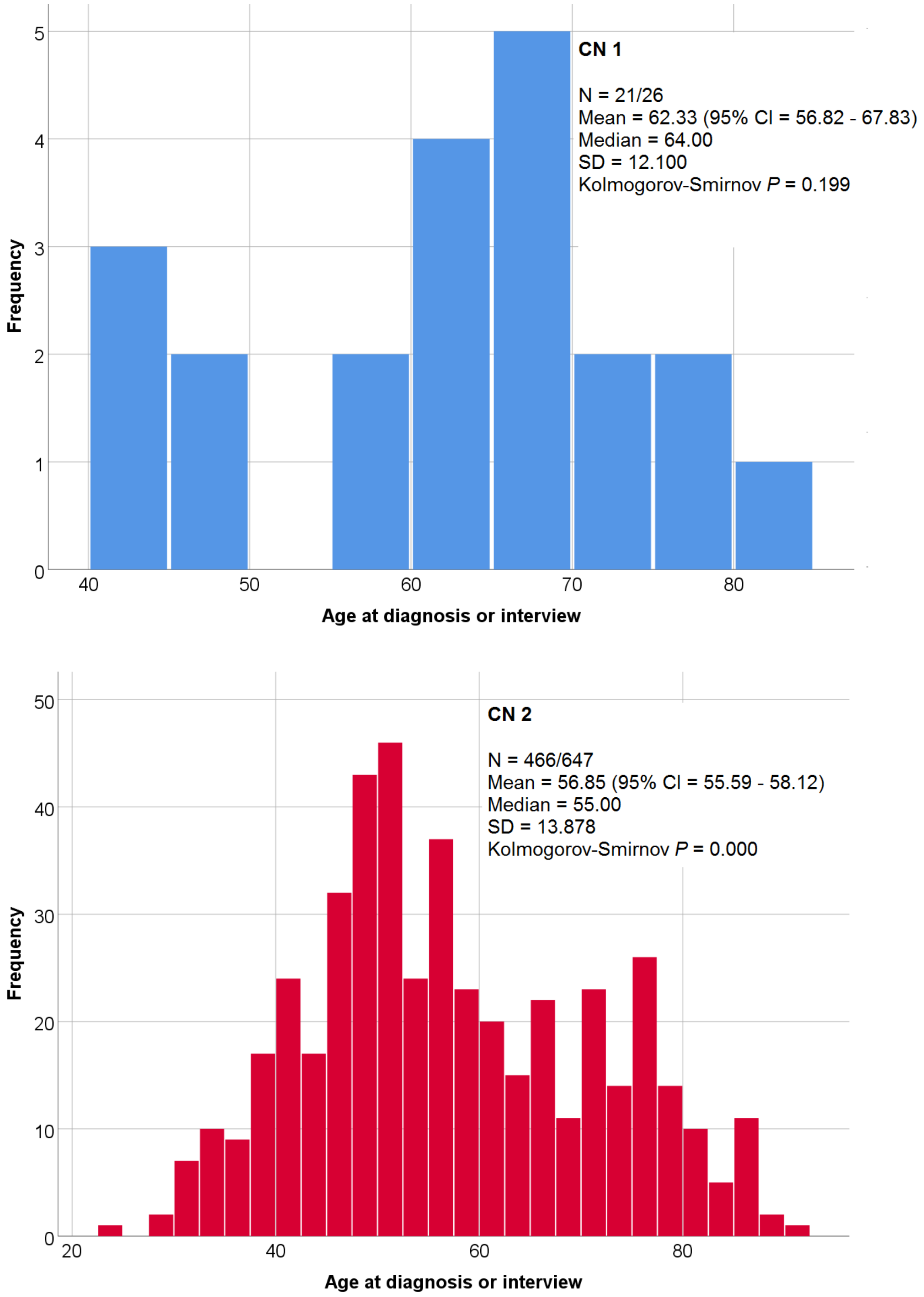


Figure 3A. Age distribution of invasive breast cancer cases and healthy controls among the single plate analyzed CN status. The frequency of participants is displayed in Y axis and the age at the time of diagnosis or interview is shown in X axis. Descriptive values are also depicted. N = 26 (all invasive cases and healthy controls having one copy of *CASC8* [blue]); N = 647 (all invasive cases and healthy controls having two copies of *CASC8* [red]). *P*-values for association of the germline CNV in *CASC8* and age of study participants are 0.010 (ages of ≤ 39, 40–49, 50–59, 60–69 and 70 ≥, evaluated by the Fisher's exact test) and 0.010 (ages of < 60 and 60 ≥, evaluated by the Pearson's chi-squared test).

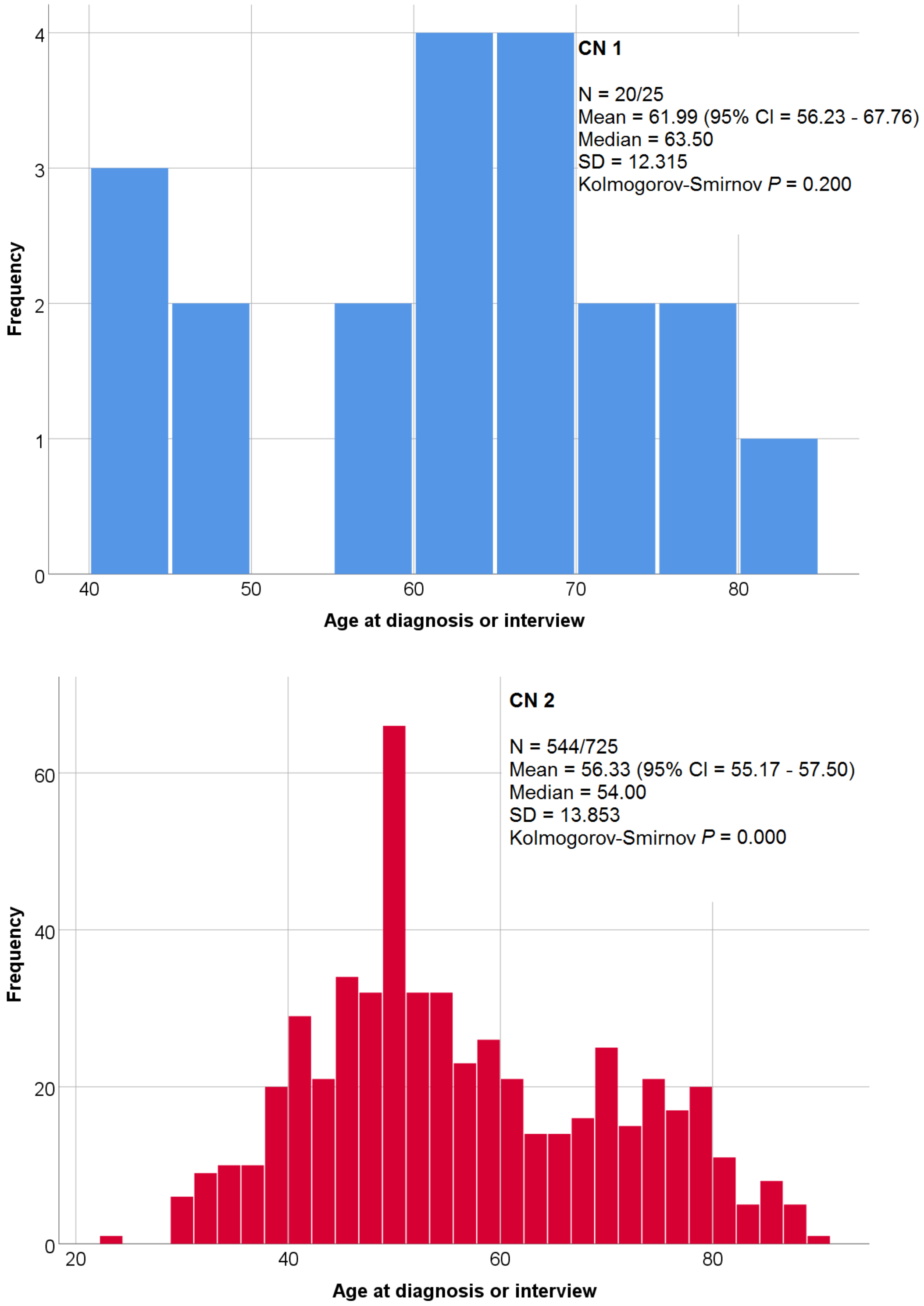


Figure 3B. Age distribution of invasive breast cancer cases and healthy controls among the multiplate analyzed CN status. The frequency of participants is displayed in Y axis and the age at the time of diagnosis or interview is shown in X axis. Descriptive values are also depicted. N = 25 (all invasive cases and healthy controls having one copy of *CASC8* [blue]); N = 725 (all invasive cases and healthy controls having two copies of *CASC8* [red]). *P*-values for association of the germline CNV in *CASC8* and age of study participants are 0.015 (ages of ≤ 39, 40–49, 50–59, 60–69 and 70 ≥, evaluated by the Fisher’s exact test) and 0.009 (ages of < 60 and 60 ≥, evaluated by the Pearson’s chi-squared test).

The CN deletion in CASC8 had no effect on BCSS or RFS times

The Kaplan–Meier plots were created based on the *CASC8* CN status to evaluate the differences in BCSS and RFS times between invasive breast cancer cases having one gene copy versus diploid gene copy, the significance of differences assessed by the log-rank Mantel–Cox test. Using the single plate analysis, 5 out of 19 invasive cases with CN deletion (26.3%) and 104 out of 373 (2 metastasized excluded) with diploid CN (27.9%) had died from breast cancer with estimated mean survival years 18.38 (95% CI = 14.49–22.27) and 19.24 (18.32–20.15), respectively (Fig. 4A). Corresponding frequencies for getting relapse/new breast cancer were 6 out of 19 (31.6%) and 147 out of 373 (39.4%) with mean disease-free years 17.63 (13.57–21.70) and 16.65 (15.61–17.69, Fig. 5A). Using the multiplate analysis, 5 out of 18 one copy cases (27.8%) and 113 out of 406 (4 metastasized excluded) diploid copy cases (27.8%) had died from breast cancer (mean survival years 18.06 [13.98–22.14] and 19.28 [18.40–20.15], Fig. 4B) whereas 5 out of 18 (27.8%) and 160 out of 406 (39.4%) had relapse/new breast cancer after 17.77 (13.46–22.07) and 16.69 (15.69–17.68) mean disease-free years (Fig. 5B). The CN deletion in *CASC8* did not show statistically significant association with BCSS or RFS times, the estimated single plate and multiplate log-rank *P*-values being 0.842 and 0.982 for BCSS and 0.436 and 0.331 for RFS, respectively.

The CN deletion in CASC8 had no relevance to breast cancer prognosis

The Cox proportional hazards model (Table 3) was used to estimate covariate-adjusted survival of the invasive breast cancer cases having one gene copy versus diploid gene copy of *CASC8*. Considering the single plate analysis, 94 out of 342 cases who died of breast cancer and with known CN (27.5%) and 134 out of 343 relapsed/new breast cancer diagnosed cases with known CN (39.1%) had information on age at breast cancer diagnosis, as well as the histology, size, grade, lymph node status and hormone receptor status (ER, PR and HER2) of breast tumor. Analyzing by the multiplate, the clinico-pathological data was available from 103 out of 371 cases who died of breast cancer (27.8%) and 146 out of 372 cases who had relapse/new breast cancer (39.2%). The CN deletion in *CASC8* did not show to associate with covariate-adjusted BCSS (*P* = 0.526 and 0.447) or RFS (*P* = 0.957 and 0.769) using the single and the multiplate analyses, respectively. Invasive cases with involved regional lymph nodes had 3.24/3.01-fold increased risk to die from breast cancer (*P* = 0.000; 95% CIs for the single and multiplate HRs 2.11–4.98 and 2.01–4.53) and 2.40/2.20-fold increased risk to get relapse or new breast cancer (*P* = 0.000; CI = 1.70–3.37 and 1.58–3.05). HER2-positivity also increased risks to short BCSS

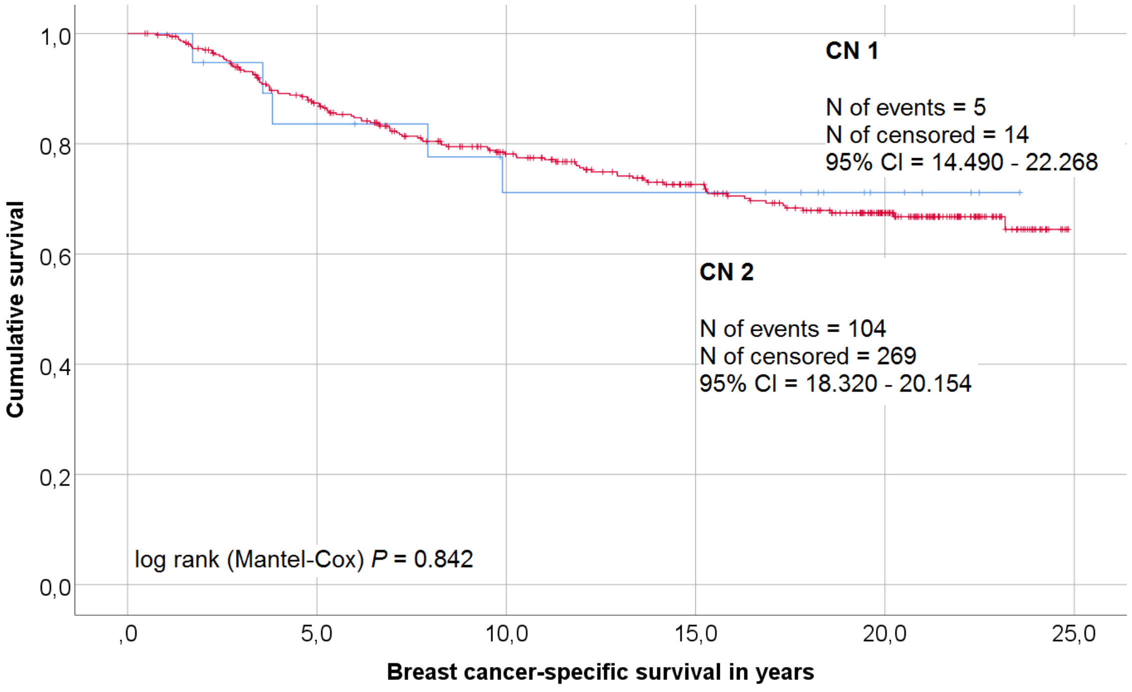
(HR = 2.01 [1.23–3.27, *P* = 0.005] and HR = 2.09 [1.32–3.30, *P* = 0.002]) and RFS times (HR = 1.56 [1.02–2.39, *P* = 0.040], only significant for multiplate analyzed cases). As summarized in Table 3, age at breast cancer diagnosis and tumor histology, size, grade and ER/PR status did not significantly associate with poor BCSS or RFS times using either CN analysis method.

Table 3. Effect of the germline CNV in *CASC8* on breast cancer prognosis

Covariates	BCSS			RFS		
	N[a]	HR (95% CI)	*P* -value[c]	N[b]	HR (95% CI)	*P*-value[c]
[1] = Single plate analysis	94/342			134/343		
[2] = Multiplate analysis	103/371			146/372		
Lymph node status						
N0 vs. N1+N2+N3	[1]	3.243 (2.111–4.979)	**0.000**	[1]	2.395 (1.701–3.374)	**0.000**
	[2]	3.013 (2.005–4.526)	**0.000**	[2]	2.197 (1.582–3.051)	**0.000**
HER2 status						
positive vs. negative	[1]	2.005 (1.230–3.267)	**0.005**	[1]	–	0.128
	[2]	2.087 (1.318–3.304)	**0.002**	[2]	1.564 (1.021–2.394)	**0.040**
Tumor size						
T1 vs. T2 vs. T3+T4	[1]	–	0.088	[1]	–	0.173
	[2]	–	0.085	[2]	–	0.222
Tumor grade						
1 vs. 2 vs. 3+4	[1]	–	0.117	[1]	–	0.292
	[2]	–	0.165	[2]	–	0.586
Age at diagnosis						
≤39, 40-49, 50-59, 60-69, 70≥	[1]	–	0.178	[1]	–	0.100
	[2]	–	0.366	[2]	–	0.130
CN status for CASC8						
deletion vs. diploid	[1]	–	0.526	[1]	–	0.957
	[2]	–	0.447	[2]	–	0.769
ER status						
positive vs. negative	[1]	–	0.573	[1]	–	0.435
	[2]	–	0.572	[2]	–	0.927
Tumor histology						
ductal vs. lobular vs. other ca	[1]	–	0.837	[1]	–	0.910
	[2]	–	0.902	[2]	–	0.970
PR status						
positive vs. negative	[1]	–	0.856	[1]	–	0.988
	[2]	–	0.816	[2]	–	0.571

a: N = 342 and 371 (all invasive cases with information on *CASC8* CN [deletion or diploid], as well as BCSS)
b: N = 343 and 372 (all invasive cases with information on *CASC8* CN [deletion or diploid], as well as RFS)
c: *P*-values were calculated using the multivariate Cox proportional hazard model stratified by major covariates.

A



B

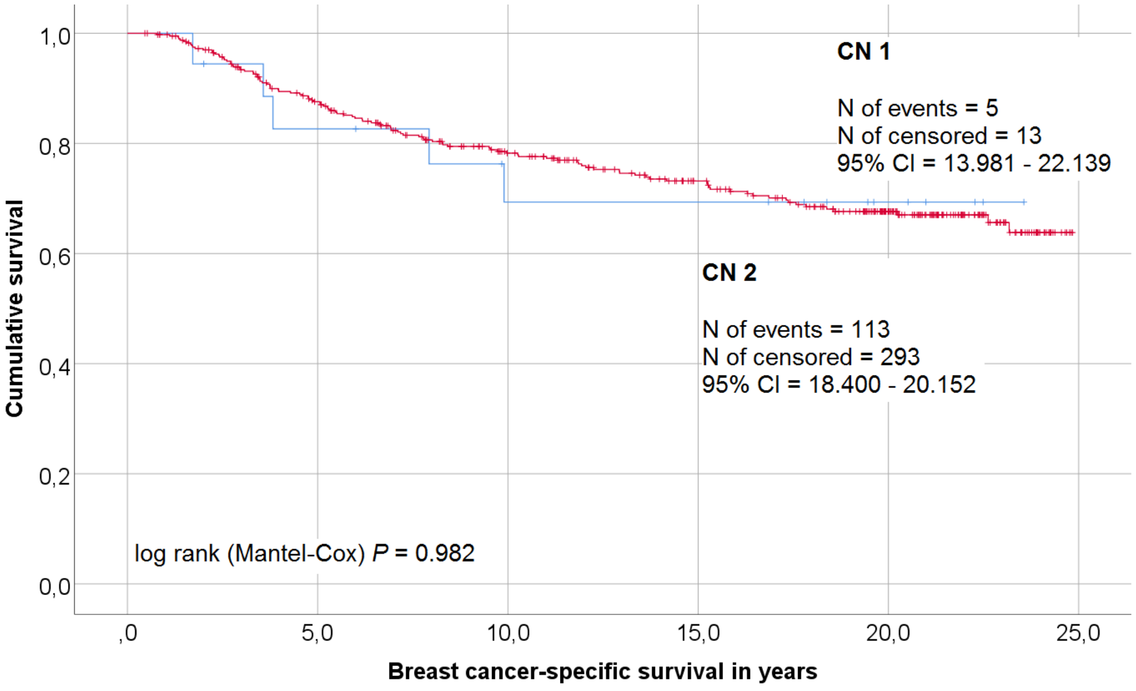


Figure 4. Kaplan–Meier plots for BCSS of invasive cases having the different CN status of *CASC8*. The cumulative survival with different CNs (deletion/one copy of *CASC8* [blue]; diploid/two copies of *CASC8* [red]) is indicated in Y axis and the BCSS in years is shown in X axis. Events = have died from breast cancer whereas censored (+) cases = are either alive or dead from other causes. 95% CI for BCSS rate and log-rank (Mantel–Cox) *P*-value for significance between the curves are also depicted. (**A**) BCSS curve from single plate analysis. (**B**) BCSS curve from multiplate analysis.

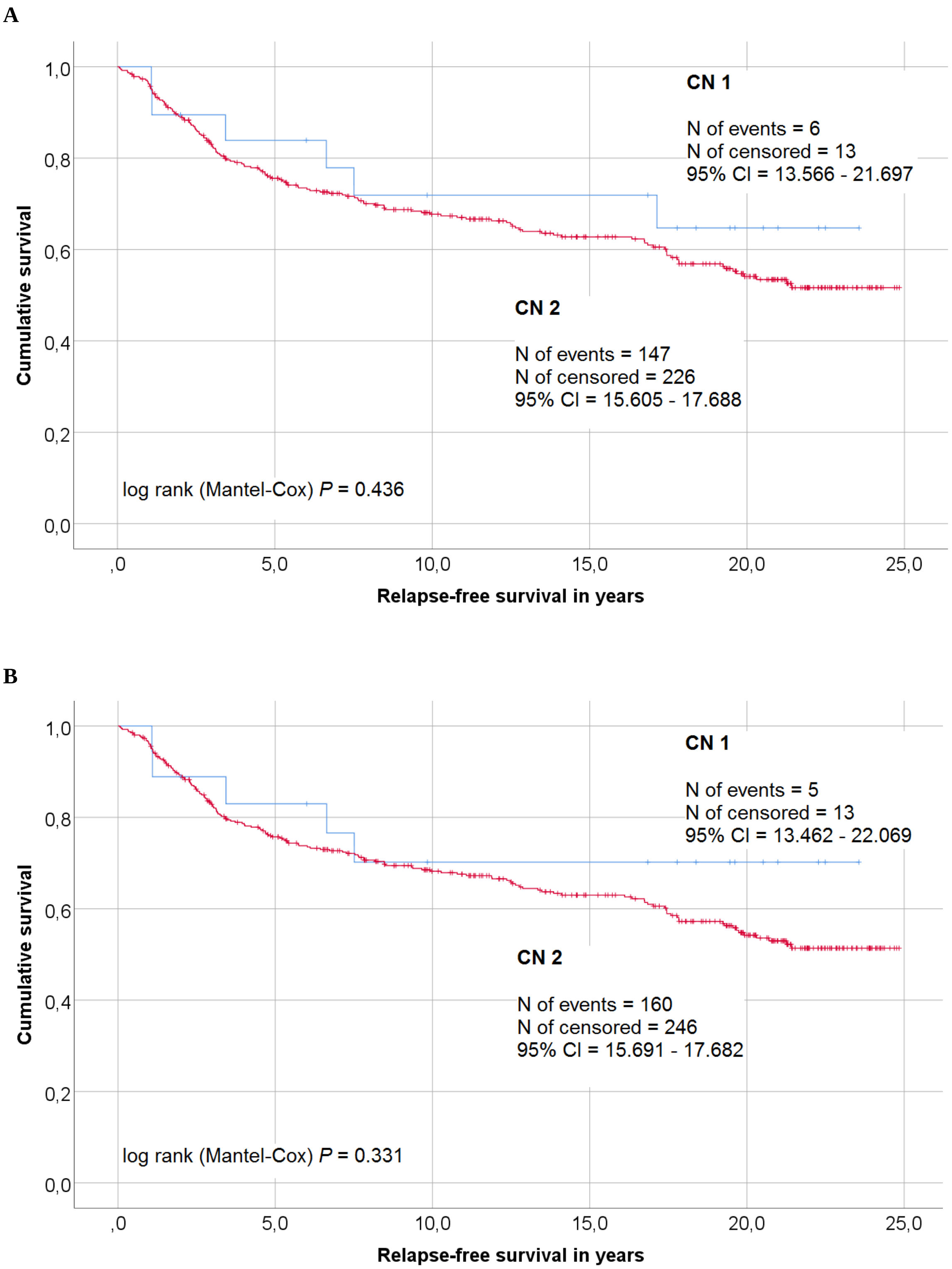


Figure 5. Kaplan–Meier plots for RFS of invasive breast cancer cases having the different CN status of *CASC8*. The cumulative survival with different CNs (deletion/one copy of *CASC8* [blue]; diploid/two copies of *CASC8* [red]) is indicated in Y axis and the RFS in years is shown in X axis. Events = relapse or diagnosed with new breast cancer whereas censored (+) cases = have not. 95% CI for RFS rate and log-rank (Mantel–Cox) *P*-value for significance between the curves are also depicted. (**A**) RFS curve from single plate analysis. (**B**) RFS curve from multiplate analysis.

Discussion

Breast cancer is the most common diagnosed cancer and cause of cancer mortality in women worldwide (Bray *et al*, 2018). One of the most important challenges in managing patients with breast cancer is that it is not a single disease, but a group of intrinsic subtypes showing distinct clinical, histological and genetic characteristics and thus, distinct clinical outcomes (Perou *et al*, 2000; Rampaul *et al*, 2001; Sørlie *et al*, 2001; Dawson *et al*, 2013). Genetic factors have a key role in the aetiology of familial and non-familial breast cancers. However, high-penetrance mutations and low-to-intermediate risk-associated variants discovered to date can only explain half of the genetic susceptibility to this malignancy. (Ghoussaini *et al*, 2013; Kumaran *et al*, 2017) A class of inherited structural variations (CNVs) in the germline DNA are recently thought to explain some of this missing heritability but however, only few risk-associated CNVs have been found so far and fewer with prognostic significance (Kumaran *et al*, 2017). The relevance of germline CNV in lncRNA gene *CASC8* to breast cancer risk and prognosis was investigated at the first time in this population-representative study involving breast cancer cases and healthy controls of Northern Savonia/Eastern Finland origin.

Using 813 gDNA samples from population-based breast cancer cases and healthy controls for CN analyses, one gene copy of *CASC8* was identified in 24 invasive cases, 2 *in situ* cases and 2 controls using the single plate analysis but in 1 invasive case less using the multiplate analysis. Respectively, diploid gene copy was found in 554 invasive cases, 47 *in situ* cases and 93 controls using the single plate analysis whereas in 589 invasive cases, 50 *in situ* cases and 136 controls using the multiplate analysis. Both CN analysis methods were used since it was not known which would be better. The highest quality CN data was shown to be generated using the single plate analysis where experimental ΔC_T variation was low and CN value ranges were set more accurately, even by removal of an outlier C_T technical replicate of the associated gDNA sample from CN analysis (e.g. insufficient mixing and/or imprecise pipetting of the replicate). The multiplate analysis was shown to be useful to determine the number of *CASC8* copies in poor quality gDNA samples showing intermediate CNs that could not otherwise be analyzed from single plates (N = 80). However, additional plate-to-plate ΔC_T variation due to differences in used DNA extraction methods (i.e. purity of samples), gDNA amounts and reagent lots between single plates could affect the quality of the multiplate analyzed CN results (Applied Biosystems CopyCaller® Software v2.0, 2011). These methodological aspects should be considered in future multiplate CN analyses, for example by analyzing samples prepared by different DNA extraction methods or reagent lots separately.

Keeping in mind the strengths and weaknesses of the different CN analysis methods, the CN deletion in *CASC8* did not significantly differ between invasive breast cancer cases and healthy controls analyzed by either method, even though cases with *in situ* breast carcinoma were included in the analyses. This could be due to a lower number of controls compared to cases in both CN analyses, as only one third of all control samples included in the original KBCP was available. Thus, it would be reasonable to replicate the results after acquiring additional age and geographically matched control samples from, for example, the Biobank of Eastern Finland (ita-suomenbiopankki.fi/en/researchers/). The frequencies of *CASC8* deletion were 3.9% for the single plate and 3.4% for the multiplate analyses, meaning that it is not a rare genetic variation (< 1%) previously reported to confer high susceptibilities for familial and early-onset breast cancer development (Pylkäs *et al*, 2012; Masson *et al*, 2014) or a common genetic variation (> 5%) previously shown to associate with low familial (Frank *et al*, 2007) and non-familial (Xuan *et al*, 2013; Kumaran *et al*, 2017) breast cancer risks. Moreover, a rare duplication in *CASC8* (0.1%) was found by the multiplate analysis, but because it was associated with intermediate CN suggesting poor sample quality and/or experimental ΔC_T variation in the single plate analysis, and plate-to-plate ΔC_T variation could exist in the multiplate analysis (Applied Biosystems CopyCaller® Software v2.0, 2011), this duplication should be validated using another approach, such as comparative genomic hybridization or sequencing (Zarrei *et al*, 2015).

In this first population-based study, the CN deletion in *CASC8* was shown to be significantly associated with the older age of invasive cases at the time of breast cancer diagnosis, as well as the older age of invasive cases and healthy controls. The associations were not influenced by used CN analysis method or age group (10-year age category and age below or equal/over 60 years). However, since none of the 50–54-year-old study participants had one gene copy of *CASC8* compared to many of the same age who had a diploid CN, previous results may be false positives. Notably, the ILRS project including 5 invasive cases with CN deletion lacked information on ages at breast cancer diagnosis. Thus, more 50–54-year-old study participants and missing information on ages of invasive cases and healthy controls are needed to confirm the statistically significant results.

The CN deletion in *CASC8* was not found to be significantly associated with the age of controls at interview using either CN analysis method. This could be due to a low number of controls in both CN analyses and thus, acquiring additional control samples may affect the obtained results. There was also no evidence for an association between given *CASC8* structural variant and any of tumor characteristics (histology, size, stage, grade, lymph node status, distant metastasis, local/distant relapse and ER/PR/HER2 status). However, there were changes in the assessment of clinico-pathological data across timely different main projects and the tumor characteristics were missing from many cases with known CN. Thus, the non-significant results from germline CNV associations with tumor characteristics analyzed by either CN method must be questioned.

Since common SNPs associated with low-penetrance breast cancer susceptibility have little or no effect on the survival of breast cancer patients (Fasching *et al*, 2012) and structural CNVs comprise a higher region of the genome than SNPs (Zarrei *et al*, 2015; Kumaran *et al*, 2017), the prognostic significance of the germline CNV in *CASC8* among genetically isolated breast cancer patients (Hartikainen *et al*, 2005) was worthwhile to investigate. Even though germline CNVs have recently been recognized as prognostic indicators for breast (Kumaran *et al*, 2017), colorectal, (Werdyani *et al*, 2017) and prostate (Laitinen *et al*, 2016) cancers and multiple SNPs located in the *CASC8* gene as conferring risks for all these cancers (Amundadottir *et al*, 2006; Zanke *et al*, 2007; Shi *et al*, 2016), this study found no evidence of an association between different CN status of *CASC8* and breast cancer mortality or relapse analyzed by either CN method. However, only non-metastatic KBCP cases at the time of breast cancer diagnosis were used because the later ILRS project lacked information on BCSS and RFS times due to a short follow-up. Thus, the results should be validated after longer follow-up of the cases included in the ILRS project.

After stratification by major clinico-pathological covariates (age at breast cancer diagnosis, as well as the histology, size, grade, lymph node status and ER/PR/HER2 status of breast tumor), there was a clear evidence that the germline CNV in *CASC8* does not play a major role in breast cancer survival in this KBCP dataset of invasive cases from the early 1990s: HER2-positivity was shown to increase risks to breast cancer mortality and relapse by a 2-fold since no targeted therapy against the HER2 receptor was approved in Finland before August 2000 (ema.europa.eu/en/medicines/human/EPAR/herceptin). The humanized monoclonal antibody trastuzumab (Herceptin®) has significantly prolonged the survival of patients with HER2-

positive metastatic or early breast cancer after its approval. It targets to the extracellular domain of HER2 and thus prevents the activation of downstream pathways related to cell proliferation, survival and apoptosis via the inhibition of dimerization of epidermal growth factor receptor family members, the internalization and degradation of HER2 receptors and the recruitment of immune cells to lyse HER2 over-expressing cells through antibody-dependent cell-mediated cytotoxicity. (Plosker & Keam, 2006) Apart from the amplification and/or over-expression of this oncogene related to the development and progression of breast cancer, invasive cases having involved regional lymph nodes had around 3-fold increased risk to die from breast cancer and around 2-fold increased risk to get relapse. These results are consistent with a large number of studies that the extent of regional lymph node involvement is one of the most significant prognostic indicators for survival in women with invasive breast carcinoma (Carter *et al*, 1989; Clayton & Hopkins, 1993; Truong *et al*, 2005) since it is related to the intrinsic invasion capability of the cancer cells (Sørlie *et al*, 2001; Dawson *et al*, 2013).

Germline CNVs in the protein-coding genes have been found to contribute to breast cancer development by disrupting the coding region of the gene (Frank *et al*, 2007; Pylkäs *et al*, 2012; Xuan *et al*, 2013) or altering the expression level of the dosage-sensitive gene in breast tissue (Kumaran *et al*, 2017). However, the functional impact of germline CNVs in the intronic and intergenic regions harboring regulatory elements and/or non-coding RNA genes on breast cancer pathogenesis are still poorly understood (Wyszynski *et al*, 2016; Kumaran *et al*, 2018). The non-protein-coding region of the genome has been shown to be highly enriched in structural CNVs, thus influencing the disease phenotype via the differences in gene expression regulation (Zarrei *et al*, 2015; Kumaran *et al*, 2018). It has been considered that the lncRNA gene *CASC8* in the gene-desert region of the chromosome 8q24.21 regulates the expression of epithelial cancer-related *proto-oncogene c-MYC* through the long-range mode of action (Ahmadiyeh *et al*, 2010; Cui *et al*, 2018). Furthermore, the expression level of that lncRNA gene has been shown to be correlated with an increased cancer risk (Cui *et al*, 2018). c-MYC is a key mediator in normal mammary gland development, as well as breast tumorigenesis and metastasis. It regulates the transcription of several genes involved in cell differentiation, growth, proliferation and apoptosis, as well as angiogenesis and stem cell fate. (Hynes & Stoelzle, 2009) Thus, the germline CNV in *CASC8* may confer breast cancer risk through aberrant c-MYC signaling that promotes mammary epithelial transformation, but the underlying mechanism is not clear to date.

This was the first case-control study to investigate the impact of germline CNV in *CASC8* on breast cancer risk and prognosis among relatively stable and genetically isolated population. One copy of this lncRNA gene was shown to be significantly associated with breast cancer diagnosis among older women and it was more often found in older invasive cases and healthy controls. Further studies containing more population-representative participants are needed to validate the obtained results and to elucidate the underlying molecular mechanisms and pathways associated with that structural variation. In conclusion, germline CNVs are an alternate source of genetic variation for breast cancer risk and their investigation can help us to develop more personalized strategies for the early diagnosis and treatment of this malignancy responsible for the suffering of millions of women annually.

Acknowledgements

I warmly thank my supervisors Professor PhD Arto Mannermaa and Adjunct Professor PhD Jaana Hartikainen for familiarizing me with the research of breast cancer genetics. I am deeply grateful to Early Stage Researchers MSc Jouni Kujala, MSc Stralina Eneh and MSc Suvi Laulainen for their close support and friendship in project’s ups and downs. Senior Laboratory Technician Helena Kemiläinen and Laboratory Technician Eija Myöhänen, I can’t praise you enough from helping me with practical issues. I warmly thank the other staff members of the Laboratory of Clinical Pathology and Forensic Medicine and the staff of the Biobank of Eastern Finland for their friendly support and companionship. My endlessly caring, loving and encouraging parents and siblings, thanks for sharing this Master of Science thesis journey with me. ‘Tumppi’, I will never forget those flowers and strawberries you brought to cheer me up. I am happy to share my life with a caring, supporting and understanding man like you whose every glance is a little piece of heaven. The financial support of the project provided by Finnish cancer organizations is also acknowledged.

References

Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, He HH, Brown M, Liu XS, Davis M, Caswell JL, Beckwith CA, Hills A, Macconaill L, Coetzee GA, Regan MM, Freedman ML (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* 107: 9742–9746

Allred DC (2010) Issues and updates: evaluating estrogen receptor-alpha, progesterone receptor, and HER2 in breast cancer. *Mod Pathol* 23: S52–59

Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, Jakobsdottir M, Kostic J, Magnusdottir DN, Ghosh S, Agnarsson K, Birgisdottir B, Le Roux L, Olafsdottir A, Blondal T, Andresdottir M, *et al*. (2006) A common variant associated with prostate cancer in European and African populations. *Nat Genet* 38: 652–658

Applied Biosystems CopyCaller® Software v2.0: User Guide, Part Number 4400042 Rev. C (2011) Life Technologies Corporation, Carlsbad, CA, USA

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 0: 1–31

Brierley JD, Gospodarowicz MK, Wittekind C (2017) Introduction. In *Union for International Cancer Control: TNM Classification of Malignant Tumours*, Van Eycken E (ed) pp 1–15. Chichester: Wiley-Blackwell

Caldarella A, Crocetti E, Bianchi S, Vezzosi V, Urso C, Biancalani M, Zappa M (2011) Female breast cancer status according to ER, PR and HER2 expression: a population based analysis. *Pathol Oncol Res* 17: 753–758

Carter CL, Allen C, Henson DE (1989) Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* 63: 181–187

Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, Watson M, Davies S, Bernard PS, Parker JS, Perou CM, Ellis MJ, Nielsen TO (2009) Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst* 101: 736–750

CHEK2 Breast Cancer Case-Control Consortium (2004) CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am J Hum Genet* 74: 1175–1182

Clayton F, Hopkins CL (1993) Pathologic correlates of prognosis in lymph node-positive breast carcinomas. *Cancer* 71: 1780–1790

Cui Y, Whiteman MK, Flaws JA, Langenberg P, Tkaczuk KH, Bush TL (2002) Body mass and stage of breast cancer at diagnosis. *Int J Cancer* 98: 279–283

Cui Z, Gao M, Yin Z, Yan L, Cui L (2018) Association between lncRNA CASC8 polymorphisms and the risk of cancer: a meta-analysis. *Cancer Manag Res* 10: 3141–3148

Davis BW, Gelber RD, Goldhirsch A, Hartmann WH, Locher GW, Reed R, Golouh R, Säve-Söderbergh J, Holloway L, Russell I, Rudenstam CM (1986) Prognostic significance of tumor grade in clinical trials of adjuvant therapy for breast cancer with axillary lymph node metastasis. *Cancer* 58: 2662–2670

Dawson SJ, Rueda OM, Aparicio S, Caldas C (2013) A new genome-driven integrated classification of breast cancer and its implications. *EMBO J* 32: 617–628

DeSantis CE, Bray F, Ferlay J, Lortet-Tieulent J, Anderson BO, Jemal A (2015) International variation in female breast cancer incidence and mortality rates. *Cancer Epidemiol Biomarkers Prev* 24: 1495–1506

Dossus L, Boutron-Ruault MC, Kaaks R, Gram IT, Vilier A, Fervers B, Manjer J, Tjonneland A, Olsen A, Overvad K, Chang-Claude J, Boeing H, Steffen A, Trichopoulou A, Lagiou P, Sarantopoulou M, Palli D, Berrino F, Tumino R, Vineis P, *et al.* (2014) Active and passive cigarette smoking and breast cancer risk: results from the EPIC cohort. *Int J Cancer* 134: 1871–1888

Elston CW, Ellis IO (1991) Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 19: 403–410

Elston CW, Ellis IO, Pinder SE (1999) Pathological prognostic factors in breast cancer. *Crit Rev Oncol Hematol* 31: 209–223

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, Mattila H, Bell DW, Haber DA, Grip M, Reiman M, Jukkola-Vuorinen A, Mustonen A, *et al*. (2007) A recurrent mutation in PALB2 in Finnish cancer families. *Nature* 446: 316–319

Fasching PA, Pharoah PD, Cox A, Nevanlinna H, Bojesen SE, Karn T, Broeks A, van Leeuwen FE, van't Veer LJ, Udo R, Dunning AM, Greco D, Aittomäki K, Blomqvist C, Shah M, Nordestgaard BG, Flyger H, Hopper JL, Southey MC, Apicella C, *et al*. (2012) The role of genetic breast cancer susceptibility variants as prognostic factors. *Hum Mol Genet* 21: 3926–3939

Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, Gavin A, Visser O, Bray F (2018) Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. *Eur J Cancer* 103: 356–387

Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, Kiechle-Bahat M, Bugert P, Schmutzler RK, Bartram CR, Burwinkel B (2007) Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis* 28: 1442–1445

Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, DiCioccio RA, Whittemore AS, Gayther SA, Giles GG, Guy M, Edwards SM, Morrison J, Donovan JL, Hamdy FC, Dearnaley DP, *et al*. (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* 100: 962–966

Ghoussaini M, Pharoah PDP, Easton DF (2013) Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? *Am J Pathol* 183: 1038–1051

Guo Q, Burgess S, Turman C, Bolla MK, Wang Q, Lush M, Abraham J, Aittomäki K, Andrulis IL, Apicella C, Arndt V, Barrdahl M, Benitez J, Berg CD, Blomqvist C, Bojesen SE, Bonanni B, Brand JS, Brenner H, Broeks A, *et al.* (2017) Body mass index and breast cancer survival: a Mendelian randomization analysis. *Int J Epidemiol* 46: 1814–1822

Hartikainen JM, Tuhkanen H, Kataja V, Dunning AM, Antoniou A, Smith P, Arffman A, Pirskanen M, Easton DF, Eskelinen M, Uusitupa M, Kosma VM, Mannermaa A (2005) An autosome-wide scan for linkage disequilibrium-based association in sporadic breast cancer cases in eastern Finland: three candidate regions found. *Cancer Epidemiol Biomarkers Prev* 14: 75–80

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 (2006) RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability. *Carcinogenesis* 27: 1593–1599

Henson DE, Ries L, Freedman LS, Carriaga M (1991) Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. The basis for a prognostic index. *Cancer* 68: 2142–2149

Hynes NE, Stoelzle T (2009) Key signalling nodes in mammary gland development and cancer: Myc. *Breast Cancer Res* 11: 210

Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A (2010) Germline copy number variation and cancer risk. *Curr Opin Genet Dev* 20: 282–289

Kumaran M, Cass CE, Graham K, Mackey JR, Hubaux R, Lam W, Yasui Y, Damaraju S (2017) Germline copy number variations are associated with breast cancer risk and prognosis. *Sci Rep* 7: 14621

Kumaran M, Krishnan P, Cass CE, Hubaux R, Lam W, Yasui Y, Damaraju S (2018) Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation. *Sci Rep* 8: 7529

Laitinen VH, Akinrinade O, Rantapero T, Tammela TL, Wahlfors T, Schleutker J (2016) Germline copy number variation analysis in Finnish families with hereditary prostate cancer. *Prostate* 76: 316–324

Masson AL, Talseth-Palmer BA, Evans TJ, Grice DM, Hannan GN, Scott RJ (2014) Expanding the genetic basis of copy number variation in familial breast cancer. *Hered Cancer Clin Pract* 12: 15

Palacios J, Robles-Frías MJ, Castilla MA, López-García MA, Benítez J (2008) The molecular pathology of hereditary breast cancer. *Pathobiology* 75: 85–94

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. *Nature* 406: 747–752

Plosker GL, Keam SJ (2006) Trastuzumab: a review of its use in the management of HER2-positive metastatic and early-stage breast cancer. *Drugs* 66: 449–475

Pylkäs K, Vuorela M, Otsukka M, Kallioniemi A, Jukkola-Vuorinen A, Winqvist R (2012) Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. *PLoS Genet* 8: e1002734

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, Evans DG, Eccles D; Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 39: 165–167

Rakha EA, El-Sayed ME, Lee AH, Elston CW, Grainge MJ, Hodi Z, Blamey RW, Ellis IO (2008) Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. *J Clin Oncol* 26: 3153–3158

Rampaul RS, Pinder SE, Elston CW, Ellis IO, Nottingham Breast Team (2001) Prognostic and predictive factors in primary breast cancer and their role in patient management: The Nottingham Breast Team. *Eur J Surg Oncol* 27: 229–238

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; Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* 38: 873–875

Rojas K, Stuckey A (2016) Breast cancer epidemiology and risk factors. *Clin Obstet Gynecol* 59: 651–672

Romieu I, Scoccianti C, Chajès V, de Batlle J, Biessy C, Dossus L, Baglietto L, Clavel-Chapelon F, Overvad K, Olsen A, Tjønneland A, Kaaks R, Lukanova A, Boeing H, Trichopoulou A, Lagiou P, Trichopoulos D, Palli D, Sieri S, Tumino R, *et al.* (2015) Alcohol intake and breast cancer in the European prospective investigation into cancer and nutrition. *Int J Cancer* 137: 1921–1930

Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, Evans DG, Eccles D; Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR, Rahman N (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* 38: 1239–1241

Shi J, Zhang Y, Zheng W, Michailidou K, Ghoussaini M, Bolla MK, Wang Q, Dennis J, Lush M, Milne RL, Shu XO, Beesley J, Kar S, Andrulis IL, Anton-Culver H, Arndt V, Beckmann MW, Zhao Z, Guo X, Benitez J, *et al.* (2016) Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. *Int J Cancer* 139: 1303–1317

Steindorf K, Ritte R, Eomois PP, Lukanova A, Tjonneland A, Johnsen NF, Overvad K, Østergaard JN, Clavel-Chapelon F, Fournier A, Dossus L, Teucher B, Rohrmann S, Boeing H, Wientzek A, Trichopoulou A, Karapetyan T, Trichopoulos D, Masala G, Berrino F, *et al.* (2013) Physical activity and risk of breast cancer overall and by hormone receptor status: The European Prospective Investigation into Cancer and Nutrition. *Int. J. Cancer* 132: 1667–1678

Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 98: 10869–10874

Tao Z, Shi A, Lu C, Song T, Zhang Z, Zhao J (2015) Breast cancer: epidemiology and etiology. *Cell Biochem Biophys* 72: 333–338

Truong PT, Berthelet E, Lee J, Kader HA, Olivotto IA (2005) The prognostic significance of the percentage of positive/dissected axillary lymph nodes in breast cancer recurrence and survival in patients with one to three positive axillary lymph nodes. *Cancer* 103: 2006–2014

Werdyani S, Yu Y, Skardasi G, Xu J, Shestopaloff K, Xu W, Dicks E, Green J, Parfrey P, Yilmaz YE, Savas S (2017) Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. *Cancer Med* 6: 1220–1232

Wyszynski A, Hong CC, Lam K, Michailidou K, Lytle C, Yao S, Zhang Y, Bolla MK, Wang Q, Dennis J, Hopper JL, Southey MC, Schmidt MK, Broeks A, Muir K, Lophatananon A, Fasching PA, Beckmann MW, Peto J, Dos-Santos-Silva I, *et al.* (2016) An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. *Hum Mol Genet* 25: 3863–3876

Xuan D, Li G, Cai Q, Deming-Halverson S, Shrubsole MJ, Shu XO, Kelley MC, Zheng W, Long J (2013) APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. *Carcinogenesis* 34: 2240–2243

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, *et al*. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 39: 989–994

Zarrei M, MacDonald JR, Merico D, Scherer SW (2015) A copy number variation map of the human genome. *Nat Rev Genet 16*: 172–183