Candidate SNP analyses integrated with mRNA expression and hormone levels reveal influence on mammographic density and breast cancer risk

Biong M.[1], Suderman M.[2,3,4*], Haakensen VD.[1,5*], Kulle B.[6,7], Berg PR.[8], Gram I.T.[9],Dumeaux V.[9], Ursin G.[10,11,12], Helland Å[1], H Hallett M.[2], Børresen-Dale AL[1,5], Kristensen V.N.[1,5,13]

Affiliations

[1]Department of Genetics, Institute for Cancer Research, The Norwegian Radium Hospital, Montebello 0310, Oslo, Norway

[2]Goodman Cancer Centre and McGill Centre for Bioinformatics, Montreal, Quebec, 3649 Sir William Osler, Montreal, Quebec, H3G 1Y6, Canada.

[3]Sackler Program for Epigenetics & Developmental Psychobiology at McGill University, McGill University, 3655 Promenade Sir William Osler, Montreal, Quebec, H3G 1Y6, Canada.

[4]Department of Pharmacology and Therapeutics, McGill University, 3655 Promenade Sir William Osler, Montreal, Quebec, H3G 1Y6, Canada.

[5]Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, NO-0315, Norway

[6]Epi-Gen, Institute of Clinical Medicine, Akershus University Hospital, University of Oslo Oslo, Norway

[7]Department of Biostatistics, Institute for Basic Medical Science, University of Oslo, Oslo, Norway

[8]Centre for Integrative Genetics, The Norwegian University of Life Sciences, Aas, Norway

[9]Institute of Community Medicine, Faculty of Health Sciences, University of Tromsø, Norway

[10]Cancer Registry of Norway, Oslo. Norway

[11]Department of Nutrition, School of Medicine, University of Oslo, Oslo, NO-0315, Norway

[12]Department of Preventive Medicine University of Southern California, Keck School of Medicine, Los Angeles, CA, USA

[13]Department for Clinical Molecular Biology (EpiGen), Institute for Clinical Medicine, Akershus University Hospital, University of Oslo, Oslo, NO-0315, Norway

Margarethe Biong: Margarethe.Biong@rr-research.no

Matthew Suderman: msuder@mcb.mcgill.ca

Vilde Drageset Haakensen: vilde.drageset.haakensen@rr-research.no

Åslaug Helland: Aslaug.Helland@rr-research.no

Bettina Kulle: b.k.andreassen@medisin.uio.no

Paul Berg: p.r.berg@bio.uio.no

Inger Torhild Gram: inger.gram@uit.no

Vanessa Dumeaux: vanessa.dumeaux@uit.no

Giske Ursin: giske.ursin@medisin.uio.no

Michael Hallett: hallett@mcb.mcgill.ca

Anne- Lise Børresen-Dale: a.l.borresen-dale@medisin.uio.no

Vessela Kristensen: vessela.kristensen@medisin.uio.no

Address correspondence to: Vessela N. Kristensen

Oslo University Hospital

The Norwegian Radium Hospital

Institute for Cancer Research

Department of Cancer Genetics

Montebello 0310, Oslo, Norway

Tel. * 47 22 78 13 75

E-mail: vessela.kristensen@medisin.uio.no

Introduction

Mammographic density (MD) is a well-documented risk factor for breast cancer[1-9], and thus a promising target for early detection of the disease. Based on mammographic results, the breast is given a percent density (PDEN) score ranging from 0-100% with 0 being the least dense (primarily adipose tissue). It is estimated that the risk of developing breast cancer is four to six times higher in women with an MD score of at least 75% [2,9] and that one-third of breast cancers occur in patients with PDEN scores of at least 50% [2]. MD is thus a considerably larger risk factor than traditional factors such as early menarche and nulliparity[1].

MD develops as a result of numerous life style factors[10,11] as well as genetic predisposition [12,13] and is greatly influenced by hormonal changes during a woman's life cycle, spanning from puberty through adulthood to menopause. The change in levels of circulating hormones such as estrogen and progesterone, and their receptors cause variations in the degree of proliferation and differentiation of the breast. Of note, during the postmenopausal years the breast is under the influence of either low levels of hormones derived from adipose tissue or, if hormone therapy (HT) is used, high levels of exogenous hormones. Also, results from studies of monozygotic and dizygotic twins show that MD has a strong genetic component accounting for as much as 30-60% of the variability[12,13]. Mutations and polymorphisms in the genes involved in the development of the breast are of great interest since they could potentially disrupt DNA binding sites and gene splicing or change the amino acid composition of the resulting proteins and influence MD. In fact, certain single nucleotide polymorphisms (SNPs) that affect MD have already been identified [14-17].

In this study, we have selected SNPs in genes belonging to the estrogen signaling pathway or with relevance to MD. A total of 257 SNPs in 165 genes were genotyped via the Sequenom MassARRAY platform and analyzed across a discovery (n=403) and a validation (n=51) dataset of subjects with extensive epidemiological information and serum hormone levels. We also generated mRNA expression profiles from breast tissue with density for a subset of these patients, in order to investigate the downstream effects of the identified genotypes.

Materials and methods

Study subjects

In total, the discovery and validation datasets contain 454 healthy, postmenopausal women (Table 1) (Supplementary Table 2 & 3) with associated SNP profiles, MD levels, hormone expression level, and additional clinical-epidemiological data.

In more detail, (1) the discovery dataset consists of postmenopausal Norwegian women from the Tromsø Mammography and Breast Cancer Study (TMBC) as described elsewhere [18]. In brief, the samples were collected as part of the population-based Norwegian Breast Cancer Screening Program (NBCSP) at the University Hospital of North Norway in the spring of 2001 and 2002. The study consists of postmenopausal women, aged 55-71 years residing in the municipality of Tromsø, Norway. Concomitant with mammography screening, the women were interviewed by a trained research nurse about reproductive and menstrual factors, previous history of cancer, smoking status, and the use of HT or other medications. Measurement of the hormones (estradiol, testosterone, DHAE, vitamin D (D4), prolactin) and glycoprotein (SHBG) were obtained for women not currently using HT. Of a total of 1041 postmenopausal women, 433 were selected for analysis based on the epidemiologic information available. Of these 433, 30 had more than 30% missing genotypes and were removed leaving 403 samples for analysis. All women signed an informed consent and the study was approved by the National Data Inspection Board and the Regional Committee for Medical Research Ethics.

(2)The validation dataset included samples from the Mammographic Density and Genetics (MDG) study as described previously [19]. This dataset contains healthy women (n=120), with low and high MD. The women were recruited either through the Norwegian Breast Cancer Screening Program between 2002 and 2007, or through referral to a breast diagnosis centre for a second look due to irregularities.

Women using anticoagulants, having breast implants or cancer, being pregnant or lactating were excluded. All women underwent mammography and provided information about weight, parity, HT use and family history of breast cancer. Two breast biopsies and three blood samples were collected from each woman. Serum levels of sex hormones were obtained (LH, FSH, prolactin, estradiol, progesterone, SHBG and testosterone). Of these, estrogen, LH and FSH were used to

determine menopausal status in combination with age and hormone use (Supplementary Table 4).

Of the 120 samples, we used only the samples for which we obtained genotypes (n=106). Of these, 10 were removed due to lack of clinical information, 45 because they were premenopausal, leaving 51 samples from healthy postmenopausal women for further analysis. All women provided signed informed consent. The study was approved by the local Regional Committee for Medical Research Ethics and local authorities (IRB approval number S-02036).

Mammographic classifications

The craniocaudal mammogram was digitized for participants in both the discovery and the validation cohorts. Only the left mammogram was obtained for subjects in the discovery cohort, whereas both left and right mammograms were obtained for the validation cohort. A Cobrascan CX-812 scanner at a resolution of 150 pixels per in. was used in the discovery cohort, whereas the Kodak Lumisys 85 scanner (Kodak, Rochester, New York) was used in the validation cohort. MD was assessed by GU and quantified using the University of Southern California Madena computer-based threshold method [20], while the total breast area was assessed by a research assistant trained by GU. Briefly, the method is as follows: the digitized mammogram is viewed on a screen, a reader defines the total breast area using an outlining tool. The region of interest (ROI) is then defined, excluding the pectoralis muscle, prominent veins and fibrous strands. A computer software program is used to determine the pixel value in the image ranging from 0, the darkest shade (black), to 225 the lightest shade (white), with shades of gray being intermediate values. The pixels are then tinted according to a certain threshold representing mammographic densities. The total number of pixels is estimated as well as the number of tinted pixels within the ROI. Absolute density is defined as the count of the tinted pixels within the ROI. Percent density (PDEN) is the absolute density divided by the total breast area multiplied by 100, and is the measurement used for subsequent analyses referred to as mammographic density (MD).

Gene and SNP selection

Using literature and molecular databases, we selected candidate genes and SNPs within genes that may potentially influence MD. A literature search was completed using Entrez Pubmed [21] with the following keywords: (1)“Estradiol and mammographic density“, (2)“Estradiol and ER”,

(3)“SNPs and (1)&(2)”, (4) “Mammographic density”, (5) “Breast density”, and (6)“Single nucleotide polymorphism and Estrogen/Progesterone”. In addition to the selection of genes from the resulting publication lists, we also included genes from the estradiol pathway as defined by CGAP [22] (provided by Biocarta) and iPATH^{TM} [23]. PathwayAssist/PathwayStudio®[24] (licensed software by Ariadne Genomics) was used to select additional genes from both the estrogen metabolism and ER signaling pathways. A total of 281 genes were selected through these processes. Candidate SNPs within these with a population frequency greater than 0.01 were subsequently identified using Ensembl [25], SNPper [26] and HapMap [27]. We then analyzed the file from HapMap in Haploview [28] using the De Bakkers “tagger”-test[29] for the identification of correlated SNPs or haplotypes (r^2>0.8). From the sets of correlated SNPs, the software arbitrary identifies one as a representative or tagging SNP, thus removing the need to test the other SNPs in the lists. Lastly we used SIFT [30] (Sorting Intolerant From Tolerant) developed at Fred Hutchinson Cancer Centre in Seattle[31] to reduce our SNP selection to those that might impact protein function based on the sequence homology and physical properties of amino acids. Other SNPs established in the literature to be associated with MD or breast cancer were also included. A total of 1001 SNPs were selected in the 281 genes. Supplementary Figure 1 outlines the gene and SNP selection process.

DNA isolation and genotyping

The DNA was isolated from blood collected in EDTA-tubes by phenol/chloroform extraction followed by ethanol precipitation using the Applied Biosystems Model 340A Nucleic Acid Extractor and stored in TE-buffer at 2-8˚C. Sample concentrations were measured by UV/Vis spectrophotometer (Nanodrop ND-1000) and normalized to 10ng/ul by adding ultrapure water.

All SNPs (n=1001) were annotated and a 200bp flanking sequence at each side of the SNP was extracted from SNPper in the CHIP bioinformatics tools[26] database. The final SNP list was sent to CIGENE at the Norwegian University of Life Sciences (UMB) in Aas for primer design. Genotyping of SNPs was performed using the MassARRAY system from Sequenom (San Diego, USA). The software SpectroDESIGNER v3.0 (Sequenom) was used to design PCR-primers, extension-primers and optimal multiplexes to ensure that little or no unspecific binding occurred due to similar primer design. In total, SNP-assays were successfully designed for 905 SNPs. An

initital subset of 519 SNPs was genotype of which 490 SNPs were successful. In quality control we removed SNPs not residing on autosomes (n=9) and that had a call rate of <80% (n=154) or a minor allele frequency (MAF) less than 5% (N=99), leaving 257 SNPs for statistical analysis (Supplementary Figure 2). Multiplex composition and primer sequences are available from the authors on request. All SNP genotyping was performed according to the iPLEX protocol from Sequenom [32]. For allele separation, the Sequenom MassARRAY™ Analyzer (Autoflex mass spectrometer) was used. Genotypes were assigned in real time by the MassARRAY SpectroTYPER RT v3.4 software (Sequenom) based on the mass peaks present. All results were manually inspected, using the MassARRAY TyperAnalyzer v3.3 software (Sequenom).

Hormone assays

The women in the discovery cohort had non-fasting venous blood samples drawn the day of the mammographic screening. Two 9mL citrate vials were collected for plasma extraction and after centrifugation for 15 minutes at 3000rpm the plasma was stored at -70°C. The hormone analyses were performed at the International Agency for Research on Cancer (IARC). Depending on the hormone, one of the three following methods were used: direct double antibody radioimmunoassays from Diagnostic Systems Laboratories(Webster, TX), direct radioimmunoassay from Immunotech (Marseille, France) or direct “sandwich” immunoradiometric assay from Cis Bio (Gif sur Yvette, France), described elsewhere[18].
In the validation cohort blood was drawn on SST tubes with gel, and then left for 30 minutes on the bench before centrifuging for 10 minutes on 2000 G. Serum was subsequently aliquoted and stored at -20°C. The serum hormone levels were measured with electrochemiluminescence immunoassays (ECLIA) on a Roche Modular E instrument under the supervision of one of the authors (VDH). The laboratory participates in an external quality assessment scheme entitled Labquality, and is accredited according to ISO–ES 17025.

Gene expression profiling from breast tissue

Biopsies from normal breast tissue were obtained from the validation dataset. Two breast biopsies were taken from each woman with a 14-gauge needle using ultrasound to identify dense areas, in order to avoid purely fatty biopsies. The biopsies were soaked in ethanol (for DNA extraction) and RNA later (for RNA extraction) and were stored at -20°C at Oslo University

Hospital. For one hospital the method differed in that the biopsies (92 patients) were snap frozen with liquid nitrogen and stored at -80°C.

The method for the RNA extraction and expression analysis is described in detail elsewhere [19]. Briefly, the homogenization, cell lysis and RNA extraction was performed with the RNeasy Mini protocol (Qiagen, Valencia,CA). Concentrations were determined using a NanoDrop ND-1000 spectrophotometer (Thermo scientific, Wilmington, DE). Amplification and labeling of the RNA was done using the Agilent Low RNA input Fluorescent Linear Amplification Kit Protocol. We used Agilent Human Whole Genome Oligo Microarrays, G4110A, processed by an Agilent scanner via Feature Extraction 9.1.3.1 software (all from Agilent Technologies, Santa Clara, CA). (See supplementary document 13 for gene expression pre-processing and normalization procedure)

Statistical analysis

The genotype distribution for the study population was assessed for deviation from Hardy Weinberg (HW) equilibrium using the chi-squared (x^2) test ($p \leq 0.05$). These SNPs were flagged in further analysis (Supplementary Table 12).

Genotype and phenotype association

All analyses were performed using R version 2.10.1, apart from the globaltest [33] which was performed in R 2.12.0. To investigate the association between the genotypes and various phenotypes (ie MD, hormone levels and gene expression), we used generalized linear models (GLMs) [34] under five different inheritance models: additive, dominant, codominant, overdominant and recessive. Supplementary Table 3 describes the values of the genotype variable under these inheritance models.

A one-way ANOVA was performed on the discovery dataset and adjusted for age and BMI to estimate the association between MD and each SNP without any reference to inheritance models(results not shown). The p-values from these F-tests, one per SNP, were corrected for multiple testing by calculating the false discovery rate (FDR).

Expression and hormone analysis

To discover factors that potentially mediate the effects of SNPs on MD in healthy women, we analyzed microarray expression data from normal breast tissue in the validation dataset as well as serum hormone levels from both cohorts. To be considered a potential mediator, we required that mRNA expression of a transcript was correlated with MD or a given hormone level (partial Pearson's correlation $p \leq 0.05$) *in cis* and associated ($p \leq 0.05$) with at least SNPs. Of the 79 healthy women in the validation dataset, 31 were postmenopausal and had genotyping and expression data available enabling their use in SNP/expression analysis with SNPassoc. Similarly, 35 of the 79 women were postmenopausal and had MD measurements available for MD/expression analysis with partial Pearson's correlation (Table 4)

Adjustments and stratification

MD decreases with age[35] and is similarly inversely correlated with BMI[36]. High BMI influences MD by way of increased areas of low density due to adipose tissue. We therefore adjusted for these effects in our model. We executed two parallel analyses, one stratifying for HT usage and the other ignoring HT usage. HT users were defined as women who were currently taking HT. We performed separate but identical analyses on all HT strata. We validated the SNPs identified in the discovery dataset associated with MD ($p \leq 0.1$).

Correlations of other variables

Partial correlation of all available variables for each cohort with adjustment for age and BMI was performed independently (Supplementary Tables 5-10).

Analysis of SNP sets

The global test [33] was used to determine associations between sets of SNPs and MD. The SNP sets were defined by using results from SNPassoc discovery analysis. For each HT strata two SNP sets were defined: one containing SNPs associated with MD at $p \leq 0.05$ and the other at $p \leq 0.01$.

Results

Gene & SNP selection

Literature search and database mining yielded 281 genes in the estradiol metabolism, ER and PR signalling pathways. The majority of genes were found by literature search (n=111) and by mining the PathwayStudio® [24]database (n=133). Other tools including CGAP and iPATH provided additional genes (n=12 and 25 respectively). Of the 281 genes we defined 725 haplotype tagging SNPs (htSNPs) using Hapmap [27] and Haploview [28]. We augmented this set with additional SNPs identified in the literature, and with SNPs located within the 281 genes that are likely to impact protein function, and, with the addition of a few fill-in SNPs (Supplementary Figure 1). The final list consisted of 1001 SNPs residing in 281 genes. Of these, an initial set of 519 SNPs (226 genes) were genotyped on the Sequenom platform. After quality control 257 SNPs (from 165 genes) remained for statistical analyses.

SNP associations

SNPs significantly associated with MD and their downstream effects on serum hormone levels and gene expression.

The discovery analyses with and without stratification by HT usage yielded in total 120 SNPs associated with MD at $p \leq 0.1$ and 77 SNPs at the standard significance level $p \leq 0.05$ under different inheritance models (results not shown). Of these 120, 31 associations were verified at $p \leq 0.1$ (no stratification n= 7; HT use n=13; non-HT use n=11, Table 2). Three SNPs overlapped between HT strata, thus 28 unique SNPs were associated with MD across all HT strata. Ten of the 28 SNPs were significantly associated with MD at $p \leq 0.05$ in both the discovery and validation datasets and are found near genes *EPOR, UGT2B28, TBP, SELENBP1* (2SNPs), *SLC7A5, UGT2B15, CARM1, PTGER3* and *SULT2A1* (Table 2).

Multiple testing correction revealed a relatively high FDR value for the 28 SNPs validated, except within the non-HT stratification. Here three SNPs associated with MD achieved an FDR less than 0.2. SNP rs1454254 in gene *UGTB15* had the lowest FDR in both datasets (FDR=0.053) (Table 2).

Among those 28 SNPs we found two associations with hormones levels in non-HT users: one SNP (rs4986942) in *GNRHR* associated with levels of SHBG (p=0.024 and p=0.051, in the discovery and validation datasets, respectively (Table 3)). The second SNP (rs1047303) in *HDS3B1* was found associated with MD in HT users and with testosterone levels in non-HT users (p=0.054 and p=0.040, in the discovery and validation datasets, respectively (Table 3)). With respect to downstream effects of these SNPs at the gene expression level, seven of the ten SNPs associated with MD at p≤ 0.05 had significant associations with 13 transcripts *in cis* (Table 4). All seven SNPs, of which five were htSNPs, were contained in a unique gene (*EPOR, UGT2B28, TBP, SLC7A5, UGT2B15, PTGER3* and *SULT2A1*).

Sets of SNPs associated with MD

In addition to the univariate validation approach described above, we applied a multivariate approach using Globaltest [31]. The SNPsets were defined by the results from the discovery univariate analysis in terms of two p-value thresholds for association with MD (p=0.01 and p=0.05) and stratification by HT use (Supplementary Table 11). For unstratified analyses, seven SNPs near seven genes (SNPset 1) associated with MD (threshold p≤0.01) were validated with a Globaltest p-value of 0.0379 (Table 5). These seven included three SNPs from the univariate analysis. In women currently using HT, 43 SNPs near 36 genes (SNPset 2) associated with MD (threshold p≤ 0.05) were validated with a Globaltest p-value of 0.034 (Table 5). Ten of these 36 were also validated univariately. For HT non-users none of the SNPsets were validated (Table 5). Figure 1 depicts the localization of the genes harbouring the SNPs found associated with MD according to the estradiol pathway.

Discussion

We have utilized a candidate gene approach in an attempt to clarify the involvement of the estradiol pathway, through the use of SNPs, in MD. Claims about associations of estradiol with MD have been conflicting with some reporting no associations [39-41] and others reporting slight associations [42]. Here we provide evidence for an association with MD by identifying SNPs associated with MD likely to affect the activity of estradiol pathway genes.

Beyond the estradiol pathway, we also obtained results providing evidence for a causal link between MD and BC via SNPs. We showed that some of these associated SNPs were also associated with circulating levels of estradiol, testosterone and SHBG. Finally, although most of the candidate genes were selected from the estradiol pathway, we found SNPs associated with MD that resides in genes that are mainly involved other cancer-related pathways such as signalling, cell cycle including the sex hormone biosynthesis pathway.

Genes involved in signalling may affect MD:

EPOR plays a major role in signalling and upon ligand binding activates the JAK/STAT and Ras/map PI3K/Akt signalling cascades [43-45]. Several studies have implicated EPOR in breast cancer[44,46]. Shi and colleagues showed that EPO, the ligand of EPOR, is able to stimulate phenotypic changes in breast cancer cell lines, and possibly induce metastasis through cell signaling[44]. The SNP (rs318699) in *EPOR* have previously been associated with polycythemia vera (PV), a disorder where blood marrow produces excessive blood cells[47]. Our results suggest that it may be linked to the development of mammographic density.

The *EPOR* SNP was found associated with the expression of *H2AFJ, HIST2H2AC* and *PHF10*. The histone family members *H2AFJ* and *HIST2H2AC* are important in processes involving DNA transcription, repair, replication and stabilization[43]. Also central in transcription regulation is *PHF10*, a gene necessary for multipotent neural stem cell renewal and proliferation[43]. The heterozygote genotype (T/C) of the *EPOR* SNP is associated with decreased expression of *H2AFJ* and higher levels of MD, thus there is an inverse relationship between MD and the expression of *H2AFJ*. This is in support of our previous findings where we found the expression of this gene to be down-regulated in high MD [19]. *H2AFJ* has been proposed to be an oncogene in breast cancer after it was found amplified and over-expressed in breast tumours [48,49]. Similarly, the expression of *HIST2H2AC* was inversely associated with MD, and the heterozygote (T/C) was associated with low expression. To our knowledge no other study has linked this gene to MD or to BC. The heterozygote genotype of rs318699 was associated with increased expression of *PHF10* and increased MD. Banga et al. 2009 reported that PHF10 function is required for cell proliferation[50] thus we hypothesise here that *PHF10* could possibly contribute to higher MD through increased cell proliferation in the breast.

Downstream of EPOR we also identified a SNP in *AKT3* associated with MD. AKT3 plays a role in cell survival, growth factor initiated signalling (e.g. insulin), angiogenesis and tumour formation[43]. Although the verification of the *AKT3* SNP is inconclusive, we observe a strong trend towards involvement of this pathway in the development of MD.

Members of cell cycle processes may affect MD:

The SNPs in *CDC20* and *MCM5* were associated with MD in postmenopausal women currently taking HT. Such women likely have elevated estrogen levels. Estradiol(E2) has previously been found to up-regulate *MCM5*, suggesting participation in DNA stability during increased cell proliferation by E2 [51]. To our knowledge *CDC20* has not been associated with hormones previously but could share the same qualities as *MCM5*.

Hormone synthesis and clearance in relation to MD:

We found seven SNPs in genes *UGT2A1, UGT2B15, UGT2B28, SULT2A1, HSD3B1* and *HSD17B2*. These genes are all central in biosynthesis or clearance of hormones in addition to other substances of special interest is the *UGT* gene family which is involved in the removal of toxic and endogenous substances such as hormones from the human body through glucuronidation. We have previously shown that women under the influence of estrogen, i.e. premenopausal or HT users, have low expression of *UGT2Bs* and high MD[19]. Our current finding further suggests the existence of SNPs in the UGT gene family that may have influence on the development of mammographic density.

The association of the SNP in UGT2B15 (Table 4) with the expression of *SCAMP1*, *SH3PXD2A* and *THRSP* is of particular interest since expression of *SCAMP1* and *SH3PXD2A* was found positively associated with MD in women not using HT. *SCAMP1* is a carrier protein proposed to participate in endocytosis [52], and *SH3PXD2A* is required for podosome formation and is found in normal immune cells and activated endothelial cells. Podosomes degrade the extracellular matrix and promote cell invasion and are thus suggested to be important in tumor invasiveness, motility and metastasis[53]. Hypothetically the degradation process of the extracellular matrix starts already in normal breast tissue by manifesting in or giving rise to mammographically dense tissue. Lastly, expression of *THRSP* was associated with the *UGT2B15* SNP and inversely

associated with MD in women not using HT. Postmenopausal women not using HT can be presumed to have undergone a normal involution and present with breast tissue mainly consisting of type 1 lobules and adipose. *THRSP* has been shown to be expressed in adipocytes, particularly in lipomatous modules. The gene has also been found expressed in lipogenic breast cancers, which suggests a role in controlling tumor lipid metabolism[43]. The *THRSP* association with breast cancer as well as adipose tissue could be explained by Kinlaw et al. who argue that the fatty acids are not carcinogenic but act to fuel metastases[54]. The association between the expression of *THRSP* and the SNP in *UGT2B15* could be explained through their shared influence by lipids, either through lipid metabolism or sex hormone biosynthesis. It is therefore not surprising that we find the expression of *THRSP* associated with MD in these women suggestive of increased expression in breasts mainly consisting of adipose tissue and low MD.

Similarly to the UGT family, *SULT2A1* encodes a protein involved in the metabolism of drugs and endogenous compounds such as hormones and ensures clearance of these compounds and limits their availability. We find here a SNP in *SULT2A1* associated with MD in postmenopausal women not currently taking HT. To our knowledge no association between SNPs in *SULT2A1* and MD has previously been reported. The implications of SULT2A1 in breast cancer mainly involves the regulation of hormone levels present in tumor tissue[55,56]. Upon analysis of expression data we found this SNP associated with the expression of *RSP26* and *NTN5* of which neither has been previously implicated in MD or BC. While little is known about ribosomal protein *RPS26*, it is known that *NTN5* codes for netrin-1-like protein which is involved in axon guidance. Netrins may also act as growth factors by encouraging cell growth activities in target cells. Specifically, netrin1 has been found to be a vascular mitogen, stimulating proliferation, inducing migration and promoting adhesion of endothelial cells and vascular smooth muscle cells[57], which could be of importance in MD.

Another regulator of hormone biosynthesis is the product of *HSD3B1* which catalyses the conversion from progneolone to progesterone. Due to its importance it has been studied in both breast cancer and mammographic density. The SNP rs1047303 has been found to be associated with MD in two independent studies [58,59]. Our findings confirm the association of the variant

type allele (Thr) with lower MD compared with the wild type allele (Asn). In addition, we found an association between this SNP and testosterone levels, a novel finding to our knowledge. A previous study found an association of the SNP with the levels of aldosterone [60]. Similarly we found a SNP in *HSD17B2* associated with MD. This gene catalyzes the conversion from estradiol to estrone thus protecting the breast tissue from estradiol. The plots of the genotype distribution in the two materials show conflicting association with MD (data not shown), but based on previous studies and a good p-value in the discovery dataset, these SNPs could potentially affect MD.

Other SNPs associated with MD:

In HT unstratified analysis we identified additional associations not mentioned previously. Of special interest is the association of a SNP in *BCAR1* (breast cancer antiestrogen resistance1) with MD. *BCAR1* is involved in cellular events such as migration, survival, transformation and invasion[43,61], all of which are important in both MD and BC. In addition, high expression levels of *BCAR1* have been linked with poor relapse-free survival and overall survival, and confers a greater risk of resistance to antiestrogen therapy such as tamoxifen [62]. This gene is located on 16q which has been implicated in many cancers including breast cancer[63,64]. We find that being homozygote for the T allele is associated with increased MD, which is to our knowledge a novel finding.

Among the women not taking hormone therapy we find SNPs with validated association with MD in the genes *MMP2* and *NOS2A (UGTB15* and *SULT2A1* previously*)*. Of special interest is *MMP2* (matrix metalloproteinase 2) which belongs to a family involved in programmed degradation of extracellular matrix during reproduction, tissue remodelling, arthritis and metastasis[65]. In particular the enzyme encoded by this gene is involved in angiogenesis, tissue repair, tumour invasion and inflammation[43], cancer cell growth, differentiation, apoptosis , migration and immune surveillance [66]. In addition it degrades type IV collagen, which is the major component in basement membranes that underline the epithelium [65]. Associations of MMPs and cancers are numerous and includes cancers of the lung, head and neck, colorectal and hormone related cancers such as, prostate, breast ,ovary and endometrium [67]. Associations of single nucleotide polymorphisms inn MMPs with cancer have also been reported[68], but for

breast cancer the results are less conclusive[69-71]. We identify here a SNP (rs2192852) in *MMP2* associated with MD in postmenopausal women not taking HT. This is an interesting gene in terms of mammographic density due to its metaplastic ability and could potentially connect MD with BC, and thus the critical point of transition from healthy to a malignant state.

Globaltest

With the use of Globaltest we identified and validated the association of two SNPsets with MD. This result indicates that the SNPs may hold more potential in a set of SNPs than as single markers.

Limitations and strengths:

These findings support the involvement of genetic variation in genes belonging to the estradiol pathway in a premalignant process, with origin in dense mammographic tissue. There are however limitations to this study. While the discovery set is, to our knowledge, a pure control dataset with a homogenous group of women, the verification dataset is more heterogeneous, including samples from women who have had a suspicious mammogram and can be therefore considered at increased risk of developing breast cancer.

The stratification of the women into HT non-users, never users and past users may also not be optimal since the past user group may include women who had recently stopped using HT. However, we are confident that this might be true for only a few women and that the majority of the HT non-users are representative for women not under the influence of HT.

Our study considered only a few hundred candidate SNPs thus other SNPs associated with MD are missed. However, the SNPs selected were mainly htSNPs with one in linkage disequilibrium (LD) with as many as 65 SNPs, thus it could be said that we tested LD blocks as opposed to single SNP for association with MD, the causative marker(s) may therefore be in LD to the SNPs found here. Of the 257 SNPs, 89 are on the 660GWAS Illumina array. Of the 28 reported with associations here, 11 are on the 660 Illumina array. None of these 11 SNPs were found significantly associated with MD in a 660GWAS study by Lindstrøm *et al.*[72]. SNPs with lower penetrance need a greater sample size due to the stringent significance level applied when testing

100K - 1M SNPs. A candidate gene approach with association to mediators and proxymal phenotypes such as mRNA and protein expression decreases this threshold and increases the chance of detecting aberrations with low risk. We did not apply bonferroni correction in the discovery SNP/MD association analysis in SNPassoc. In our study, however, these SNPs were retained and validated in an additional cohort. Agreement of several of our findings with previous studies suggests that they should be further investigated in larger sample populations.

Conclusion

We find here associations of SNPs within genes involved in the metabolism and regulation of estrogen, a finding that strengthens the existing hypothesis of hormone dependency in MD. We have provided evidence that some of these SNPs may exert an effect on MD via changes in expression levels of certain genes, Further functional studies are required to determine the exact effects of these or other SNPs in LD.

Abbreviations

MD: mammographic density, BC:breast cancer, HT: Hormone Therapy, BMI: body mass index, SNP: single nucleotide polymorphism

Competing interests

The authors declare that they have no competing interests.

Authors contributions

The study was supported by grants to VNK from the Norwegian cancer society and the South Eastern Norway Regional Health Authority. MB was supported by grants from the Norwegian Research Council and the South Eastern Norway Regional Health Authority. IT is responsible for collecting TMBC data. ÅH is responsible for collecting MDG data. MB, PRB performed the genotyping lab work. VDH performed the expression array lab work and data processing, assisted in MDG data collection and estimation of mammographic density. GU is responsible for the mammographic density data for both cohorts used.VHD assisted in measuring the serum hormones in MDG samples. MB and MS and BK performed the statistical analysis. MB, MS, and VNK, VD and MH interpreted the results, and MB and MS wrote the paper. VD critically revised and commented on the manuscript. MH provided statistical advice and critically revised and commented on the manuscript. All authors were involved in reviewing the report.

Reference List

1. Harvey JA, Bovbjerg VE: **Quantitative assessment of mammographic breast density: relationship with breast cancer risk.** *Radiology* 2004, **230:** 29-41.

2. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ *et al*.: **Mammographic breast density as an intermediate phenotype for breast cancer.** *Lancet Oncol* 2005, **6:** 798-808.

3. Byrne C, Schairer C, Wolfe J, Parekh N, Salane M, Brinton LA *et al*.: **Mammographic features and breast cancer risk: effects with time, age, and menopause status.** *J Natl Cancer Inst* 1995, **87:** 1622-1629.

4. Byrne C, Schairer C, Brinton LA, Wolfe J, Parekh N, Salane M *et al*.: **Effects of mammographic density and benign breast disease on breast cancer risk (United States).** *Cancer Causes Control* 2001, **12:** 103-110.

5. Byrne C: **Studying mammographic density: implications for understanding breast cancer.** *J Natl Cancer Inst* 1997, **89:** 531-533.

6. Ursin G, Hovanessian-Larsen L, Parisky YR, Pike MC, Wu AH: **Greatly increased occurrence of breast cancers in areas of mammographically dense tissue.** *Breast Cancer Res* 2005, **7:** R605-R608.

7. Russo J, Mailo D, Hu YF, Balogh G, Sheriff F, Russo IH: **Breast differentiation and its implication in cancer prevention.** *Clin Cancer Res* 2005, **11:** 931s-936s.

8. Ziv E, Shepherd J, Smith-Bindman R, Kerlikowske K: **Mammographic breast density and family history of breast cancer.** *J Natl Cancer Inst* 2003, **95:** 556-558.

9. McCormack VA, dos SS, I: **Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis.** *Cancer Epidemiol Biomarkers Prev* 2006, **15:** 1159-1169.

10. Oza AM, Boyd NF: **Mammographic parenchymal patterns: a marker of breast cancer risk.** *Epidemiol Rev* 1993, **15:** 196-208.

11. Vachon CM, Kuni CC, Anderson K, Anderson VE, Sellers TA: **Association of mammographically defined percent breast density with epidemiologic risk factors for breast cancer (United States).** *Cancer Causes Control* 2000, **11:** 653-662.

12. Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR *et al*.: **Heritability of mammographic density, a risk factor for breast cancer.** *N Engl J Med* 2002, **%19;347:** 886-894.

13. Ursin G, Lillie EO, Lee E, Cockburn M, Schork NJ, Cozen W *et al*.: **The relative importance of genetics and environment on mammographic density.** *Cancer Epidemiol Biomarkers Prev* 2009, **18:** 102-112.

14. Haiman CA, Hankinson SE, De V, I, Guillemette C, Ishibe N, Hunter DJ *et al*.: **Polymorphisms in steroid hormone pathway genes and mammographic density.** *Breast Cancer Res Treat* 2003, **77:** 27-36.

15. Lai JH, Vesprini D, Zhang W, Yaffe MJ, Pollak M, Narod SA: **A polymorphic locus in the promoter region of the IGFBP3 gene is related to mammographic breast density.** *Cancer Epidemiol Biomarkers Prev* 2004, **13:** 573-582.

16. van Duijnhoven FJ, Bezemer ID, Peeters PH, Roest M, Uitterlinden AG, Grobbee DE *et al*.: **Polymorphisms in the estrogen receptor alpha gene and mammographic density.** *Cancer Epidemiol Biomarkers Prev* 2005, **14:** 2655-2660.

17. Mulhall C, Hegele RA, Cao H, Tritchler D, Yaffe M, Boyd NF: **Pituitary growth hormone and growth hormone-releasing hormone receptor genes and associations with mammographic measures and serum growth hormone.** *Cancer Epidemiol Biomarkers Prev* 2005, **14:** 2648-2654.

18. Bremnes Y, Ursin G, Bjurstam N, Rinaldi S, Kaaks R, Gram IT: **Endogenous sex hormones, prolactin and mammographic density in postmenopausal Norwegian women.** *Int J Cancer* 2007, **121:** 2506-2511.

19. Haakensen VD, Biong M, Lingjaerde OC, Holmen MM, Frantzen JO, Chen Y *et al*.: **Expression levels of uridine 5'-diphospho-glucuronosyltransferase genes in breast tissue from healthy women are associated with mammographic density.** *Breast Cancer Res* 2010, **12:** R65.

20. Ursin G, Astrahan MA, Salane M, Parisky YR, Pearce JG, Daniels JR *et al*.: **The detection of changes in mammographic densities.** *Cancer Epidemiol Biomarkers Prev* 1998, **7:** 43-47.

21. Pubmed. 2006.
Ref Type: Generic

22. CGAP. 2006.
Ref Type: Generic

23. iPATH. 2006.
Ref Type: Generic

24. Pathway Assist/Studio. 2006.
Ref Type: Computer Program

25. Ensemble. Ensemble . 1-8-2005.
Ref Type: Electronic Citation

26. CHIP. 2006.
Ref Type: Generic

27. Hapmap. 2006.
Ref Type: Generic

28. Haploview. 2006.
Ref Type: Generic

29. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies.** *Nat Genet* 2005, **37:** 1217-1223.

30. SIFT. J . 1-8-2011.
Ref Type: Computer Program

31. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res* 2001, **11:** 863-874.

32. Sequenom. 2006.
Ref Type: Generic

33. Goeman JJ, van de Geer SA, de KF, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20:** 93-99.

34. Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X *et al*.: **SNPassoc: an R package to perform whole genome association studies.** *Bioinformatics* 2007, **23:** 644-645.

35. Brisson J, Sadowsky NL, Twaddle JA, Morrison AS, Cole P, Merletti F: **The relation of mammographic features of the breast to breast cancer risk factors.** *Am J Epidemiol* 1982, **115:** 438-443.

36. Brisson J, Morrison AS, Kopans DB, Sadowsky NL, Kalisher L, Twaddle JA *et al*.: **Height and weight, mammographic features of breast tissue, and breast cancer risk.** *Am J Epidemiol* 1984, **119:** 371-381.

37. Huang dW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:** 44-57.

38. Huang dW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37:** 1-13.

39. Tamimi RM, Hankinson SE, Colditz GA, Byrne C: **Endogenous sex hormone levels and mammographic density among postmenopausal women.** *Cancer Epidemiol Biomarkers Prev* 2005, **14:** 2641-2647.

40. Boyd NF, Stone J, Martin LJ, Jong R, Fishell E, Yaffe M *et al*.: **The association of breast mitogens with mammographic densities.** *Br J Cancer* 2002, **87:** 876-882.

41. Sprague BL, Trentham-Dietz A, Gangnon RE, Buist DS, Burnside ES, iello Bowles EJ *et al*.: **Circulating sex hormones and mammographic breast density among postmenopausal women.** *Horm Cancer* 2011, **2:** 62-72.

42. Greendale GA, Palla SL, Ursin G, Laughlin GA, Crandall C, Pike MC *et al*.: **The association of endogenous sex steroids and sex steroid binding proteins with mammographic density: results from the Postmenopausal Estrogen/Progestin Interventions Mammographic Density Study.** *Am J Epidemiol* 2005, **162:** 826-834.

43. GeneCards. 4-1-2008.
Ref Type: Generic

44. Shi Z, Hodges VM, Dunlop EA, Percy MJ, Maxwell AP, El-Tanani M *et al*.: **Erythropoietin-induced activation of the JAK2/STAT5, PI3K/Akt, and Ras/ERK pathways promotes malignant cell behavior in a modified breast cancer cell line.** *Mol Cancer Res* 2010, **8:** 615-626.

45. Jelkmann W: **Erythropoietin after a century of research: younger than ever.** *Eur J Haematol* 2007, **78:** 183-205.

46. Pelekanou V, Notas G, Sanidas E, Tsapis A, Castanas E, Kampa M: **Testosterone membrane-initiated action in breast cancer cells: Interaction with the androgen signaling pathway and EPOR.** *Mol Oncol* 2010, **4:** 135-149.

47. Pardanani A, Fridley BL, Lasho TL, Gilliland DG, Tefferi A: **Host genetic variation contributes to phenotypic diversity in myeloproliferative disorders.** *Blood* 2008, **111:** 2785-2789.

48. Yao J, Weremowicz S, Feng B, Gentleman RC, Marks JR, Gelman R *et al*.: **Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression.** *Cancer Res* 2006, **66:** 4065-4078.

49. de Wit NJ, Rijntjes J, Diepstra JH, van Kuppevelt TH, Weidle UH, Ruiter DJ *et al*.: **Analysis of differential gene expression in human melanocytic tumour lesions by custom made oligonucleotide arrays.** *Br J Cancer* 2005, **92:** 2249-2261.

50. Banga SS, Peng L, Dasgupta T, Palejwala V, Ozer HL: **PHF10 is required for cell proliferation in normal and SV40-immortalized human fibroblast cells.** *Cytogenet Genome Res* 2009, **126:** 227-242.

51. Iso T, Futami K, Iwamoto T, Furuichi Y: **Modulation of the expression of bloom helicase by estrogenic agents.** *Biol Pharm Bull* 2007, **30:** 266-271.

52. Fernandez-Chacon R, Achiriloaie M, Janz R, Albanesi JP, Sudhof TC: **SCAMP1 function in endocytosis.** *J Biol Chem* 2000, **275:** 12752-12756.

53. Goicoechea SM, Bednarski B, Garcia-Mata R, Prentice-Dunn H, Kim HJ, Otey CA: **Palladin contributes to invasive motility in human breast cancer cells.** *Oncogene* 2009, **28:** 587-598.

54. Kinlaw WB, Quinn JL, Wells WA, Roser-Jones C, Moncur JT: **Spot 14: A marker of aggressive breast cancer and a potential therapeutic target.** *Endocrinology* 2006, **147:** 4048-4055.

55. Li K, Chandra DP, Foo T, Adams JB, McDonald D: **Steroid metabolism by human mammary carcinoma.** *Steroids* 1976, **28:** 561-574.

56. Adams JB, Chandra DP: **Dehydroepiandrosterone sulfotransferase as a possible shunt for the control of steroid metabolism in human mammary carcinoma.** *Cancer Res* 1977, **37:** 278-284.

57. Park KW, Crouse D, Lee M, Karnik SK, Sorensen LK, Murphy KJ *et al*.: **The axonal attractant Netrin-1 is an angiogenic factor.** *Proc Natl Acad Sci U S A* 2004, **101:** 16210-16215.

58. Haiman CA, Bernstein L, Berg D, Ingles SA, Salane M, Ursin G: **Genetic determinants of mammographic density.** *Breast Cancer Res* 2002, **4:** R5.

59. Stone J, Gurrin LC, Byrnes GB, Schroen CJ, Treloar SA, Padilla EJ *et al*.: **Mammographic density and candidate gene variants: a twins and sisters study.** *Cancer Epidemiol Biomarkers Prev* 2007, **16:** 1479-1484.

60. Shimodaira M, Nakayama T, Sato N, Aoi N, Sato M, Izumi Y *et al*.: **Association of HSD3B1 and HSD3B2 gene polymorphisms with essential hypertension, aldosterone level, and left ventricular structure.** *Eur J Endocrinol* 2010, **163:** 671-680.

61. OMIM. NCBI . 2008.
Ref Type: Electronic Citation

62. van der FS, Brinkman A, Look MP, Kok EM, Meijer-van Gelder ME, Klijn JG *et al*.: **Bcar1/p130Cas protein and primary breast cancer: prognosis and response to tamoxifen treatment.** *J Natl Cancer Inst* 2000, **%19;92:** 120-127.

63. Dutrillaux B, Gerbault-Seureau M, Zafrani B: **Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes.** *Cancer Genet Cytogenet* 1990, **49:** 203-217.

64. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA *et al*.: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2007, **39:** 865-869.

65. Entrez Gene. NCBI . 2008.
Ref Type: Electronic Citation

66. Egeblad M, Werb Z: **New functions for the matrix metalloproteinases in cancer progression.** *Nat Rev Cancer* 2002, **2:** 161-174.

67. O'Mara TA, Clements JA, Spurdle AB: **The use of predictive or prognostic genetic biomarkers in endometrial and other hormone-related cancers: justification for extensive candidate gene single nucleotide polymorphism studies of the matrix metalloproteinase family and their inhibitors.** *Cancer Epidemiol Biomarkers Prev* 2009, **18:** 2352-2365.

68. Chaudhary AK, Singh M, Bharti AC, Asotra K, Sundaram S, Mehrotra R: **Genetic polymorphisms of matrix metalloproteinases and their inhibitors in potentially malignant and malignant lesions of the head and neck.** *J Biomed Sci* 2010, **17:10.:** 10.

69. gado-Enciso I, Cepeda-Lopez FR, Monrroy-Guizar EA, Bautista-Lam JR, ndrade-Soto M, Jonguitud-Olguin G *et al*.: **Matrix metalloproteinase-2 promoter polymorphism is associated with breast cancer in a Mexican population.** *Gynecol Obstet Invest* 2008, **65:** 68-72.

70. Lei H, Hemminki K, Altieri A, Johansson R, Enquist K, Hallmans G *et al*.: **Promoter polymorphisms in matrix metalloproteinases and their inhibitors: few associations with breast cancer susceptibility and progression.** *Breast Cancer Res Treat* 2007, **103:** 61-69.

71. Zhou P, Du LF, Lv GQ, Yu XM, Gu YL, Li JP *et al*.: **Current evidence on the relationship between four polymorphisms in the matrix metalloproteinases (MMP) gene and breast cancer risk: a meta-analysis.** *Breast Cancer Res Treat* 2011, **127:** 813-818.

72. Lindstrom S, Vachon CM, Li J, Varghese J, Thompson D, Warren R *et al*.: **Common variants in ZNF365 are associated with both mammographic density and breast cancer risk.** *Nat Genet* 2011, **43:** 185-187.

Tables and figures:

Table 1: Overview of the study population with a short description and a list of the measured metabolites for each cohort included.

Data set	n (Total)	n (Anlysed)	Serum analyses	Description
TMBC	1041	403	IGF1, IGFBP3, IGFratio, Estradiol, Testosterone, SHBG, DHEA, D4, Prolactin.	Age: 55-71 years, postmenopausal. Negative mammograms, extensive diet questionnaire, menstrual and reproductive factors known.Collected for use in mammographic density study.
MDG	186	51	Estradiol, Testosterone, SHBG, FSH, LH , Progesterone, Prolactin.	Age: 25-69 years, premenopausal, menopausal and postmenopausal with dense breast and small cancers. Questionnaire on hormone use and reproductive factors. Collected for use in mammographic density study.
Total	**1227**	**454**		

$^{\Phi}$ The Mammography and Breast Cancer study

$^{\varphi}$ Mammographic Density and Genetics

Table 2: SNPs found significantly associated with PDEN in the discovery and verification set and their overlap according to p-value < 0.1, stratified on HT.

						Discovery				Verification					
	HT[3] stratification	SNP	Role	Gene	Inheritance model	n	MA[1]	MAF[2]	p-value	n	MA[1]	MAF[2]	p-value	p ≤ 0.05	FDR
PDEN	none	rs5842	3' UTR	BCAR1	recessive	391	T	0.47	0.0012	50	T	0.43	0.0858		0.2855
		rs1549926	Intron	CARM1	overdominant	395	C	0.28	0.0371	50	C	0.25	0.0828		0.3308
		rs318699	Promoter	EPOR	overdominant	390	T	0.41	0.0076	50	T	0.40	0.0161	*	0.3993
		rs320320	Intron	AKT3	recessive	360	C	0.20	0.0866	50	C	0.23	0.0090		0.4517
		rs4235126	Intron	UGT2B28	overdominant	364	A	0.18	0.0450	50	A	0.25	0.0429	*	0.5248
		rs2910393	Intron	SULT2A1	recessive	398	A	0.25	0.0216	50	A	0.23	0.0895		0.6179
		rs2288741	Intron	UGT2A1	dominant	394	C	0.17	0.0706	50	C	0.21	0.0412		0.7003
	HT user	rs3917117	Intron	RFC4	codominant	114	T	0.28	0.0496	18	T	0.28	0.0594		0.6465
		rs3815559	Intron (boundary)	SLC7A5	overdominant	113	G	0.19	0.0067	19	G	0.32	0.0428	*	0.6465
		rs2800953	Intron	SELENBP1	log.additive	108	C	0.32	0.0405	19	C	0.29	0.0086	*	0.7679
		rs4461565	Intron	SLC7A11	codominant	113	T	0.11	0.0294	19	T	0.11	0.0549		0.8250
		rs2072916	Intron	TBP	dominant	113	T	0.50	0.0178	19	C	0.45	0.0084	*	0.8452
		rs1047303	Coding exon	HSD3B1	overdominant	113	C	0.30	0.0343	19	C	0.32	0.0877		0.8525
		rs10788804	Intron (boundary)	SELENBP1	dominant	112	T	0.42	0.0456	19	T	0.34	0.0029	*	0.8705
		rs337887	Intron	ARSB	log.additive	109	A	0.47	0.0283	19	G	0.37	0.0568		0.8737
		rs710251	Intron (boundary)	CDC20	overdominant	115	C	0.35	0.0639	17	C	0.32	0.0483		0.8740
		rs7207463	Intron	NCOR1	dominant	114	C	0.49	0.0518	19	T	0.50	0.0146		0.8742
		rs9914387	Promoter	SPHK1	codominant	110	T	0.11	0.0719	19	T	0.08	0.0815		0.8967
		rs743815	Intron	MCM5	overdominant	114	T	0.40	0.0472	19	T	0.34	0.0793		0.9121
		rs4236	Coding exon	MGP	log.additive	115	G	0.34	0.0478	19	G	0.32	0.0950		0.9194
	HT non-user	rs1454254	Intron	UGT2B15	dominant	271	A	0.16	0.0018	31	A	0.24	0.0124	*	0.0527
		rs1549926	Intron	CARM1	overdominant	281	C	0.28	0.0147	31	C	0.26	0.0193	*	0.1541
		rs2300166	Intron	PTGER3	recessive	270	T	0.37	0.0260	30	C	0.47	0.0436	*	0.1783
		rs4986942	Coding exon	GNRHR	codominant	284	T	0.07	0.0435	31	T	0.16	0.0824		0.3336
		rs3822264	Intron	S100P	codominant	288	A	0.12	0.0916	31	A	0.16	0.0291		0.3373
		rs9934209	Intron	HSD17B2	codominant	265	C	0.43	0.0796	31	C	0.44	0.0409		0.3467
		rs2192852	Intron	MMP2	codominant	265	G	0.13	0.0854	31	G	0.16	0.0792		0.3754
		rs4462652*	Intron	NOS2A	codominant	260	A	0.26	0.0537	31	A	0.19	0.0468		0.3875
		rs763100	Intron	FBLN1	overdominant	257	A	0.17	0.0541	31	A	0.18	0.0652		0.4488
		rs4235126	Intron	UGT2B28	codominant	265	A	0.18	0.0554	31	A	0.32	0.0257		0.4775
		rs2910393	Intron	SULT2A1	recessive	284	A	0.26	0.0133	31	A	0.27	0.0289	*	0.5331

1) Minor Allele
2) Minor Allele Frequency
3) Hormone Therapy
* Merged with rs1113283

Table 3: SNPs found associated with testosterone and SHBG in the discovery and verification set and their overlap, HT non-user only.

				Discovery				Verification			
Serum hormone	SNP	Gene	Inheritance model	n	MA[1]	MAF[2]	p-value	n	MA[1]	MAF[2]	p-value
SHBG[3]	rs4986942	GNRHR	codominant	279	T	0.07	0.024	32	T	0.17	0.051
TES[4]	rs1047303	HSD3B1	codominant	262	C	0.26	0.054	32	C	0.33	0.040

1) Minor Allele
2) Minor Allele Frequency
3) Serum sex hormone binding globulin
4) Serum testosterone

Table 4: Table listing the significant SNP associations with PDEN ($p \leq 0.05$), the correlation of expression with PDEN, and the corresponding correlation coefficient. Results are stratified by HT.

						Discovery		Verification						
						P-value		P-value					coefficient	
HT[1)] Stratification	SNP	Gene	Probe ID	Gene	Inheritance model	SNP-PDEN association	n	SNP-PDEN association	n	SNP-Expression association	n	Expression-PDEN correlation	Expression-PDEN correlation coefficient	n
none	rs318699	EPOR	A_23_P204277	H2AFJ	overdominant	0.008	390	0.016	50	0.008	31	0.001	-0.526	35
			A_24_P8721	HIST2H2AC	overdominant	0.008	390	0.016	50	0.038	31	0.004	-0.472	35
			A_24_P535219	PHF10	overdominant	0.008	390	0.016	50	0.028	31	0.007	0.447	35
			A_23_P116694	RPS26	overdominant	0.008	390	0.016	50	9.18E-06	31	0.009	-0.439	35
	rs4235126	UGT2B28	A_32_P394951	NTN5	overdominant	0.045	364	0.043	50	0.002	31	0.030	0.374	35
			A_23_P167920	DLL1	overdominant	0.045	364	0.043	50	0.019	31	0.008	0.439	35
HT user	rs2072916	TBP	A_24_P225468	ANP32E	dominant	0.018	113	0.008	19	0.003	10	0.016	0.408	11
			A_24_P8721	HIST2H2AC	dominant	0.018	113	0.008	19	4.63E-05	10	0.004	-0.472	11
			A_24_P125894	PPM1F	dominant	0.018	113	0.008	19	0.001	10	0.027	-0.379	11
			A_23_P204277	H2AFJ	dominant	0.018	113	0.008	19	0.003	10	0.001	-0.526	11
			A_24_P712350	CHML	dominant	0.018	113	0.008	19	3.00E-12	10	0.008	0.443	11
			A_23_P167920	DLL1	dominant	0.018	113	0.008	19	0.020	10	0.008	0.439	11
	rs3815559	SLC7A5	A_23_P204277	H2AFJ	overdominant	0.007	113	0.043	19	0.034	10	0.001	-0.526	11
			A_23_P130780	GLTSCR1	overdominant	0.007	113	0.043	19	0.042	10	0.042	-0.353	11
HT non-user	rs1454254	UGT2B15	A_24_P42527	SCAMP1	dominant	0.002	271	0.012	31	2.94E-05	21	0.032	0.370	24
			A_23_P35456	SH3PXD2A	dominant	0.002	271	0.012	31	0.030	21	0.016	0.407	24
			A_23_P204277	H2AFJ	log.additive	0.002	271	0.017	31	0.040	21	0.001	-0.526	24
			A_23_P105212	THRSP	log.additive	0.002	271	0.017	31	0.013	21	0.037	-0.361	24
	rs2300166	PTGER3	A_24_P42527	SCAMP1	recessive	0.026	270	0.044	30	0.043	20	0.032	0.370	24
	rs2910393	SULT2A1	A_23_P116694	RPS26	recessive	0.013	284	0.029	31	0.031	21	0.009	-0.439	24
			A_32_P394951	NTN5	recessive	0.013	284	0.029	31	0.002	21	0.030	0.374	24

Table5: Globaltest results based on SNPsets defined from results in SNPassoc discovery analysis and validated in the verification dataset.

	SNPset	Global p-value		Global p-value	
HT strata	p-value threshold	discovery	samples(n)	Verification	samples(n)
none	≤ 0.01	3.07E-06	403	0.038	50
	≤ 0.05	3.01E-11	403	0.243	50
HT users	≤ 0.01	5.73E-07	115	0.376	19
	≤ 0.05	5.76E-17	115	0.034	19
HT non users	≤ 0.01	2.54E-05	288	0.4	31
	≤ 0.05	1.75E-13	288	0.43	31

PathDesignerEstrogen-DependentBreastCancer_modifie