From Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden

# GENETIC DETERMINANTS OF BREAST CANCER RISK

Jingmei Li

Stockholm 2011

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet. Printed by Larserics Digital AB.

ISBN [978-91-7457-179-0]

For Mommy

In my dreams, I seek the solace that only he can give –
A shelter from the storm
A mad man's heaven
A soul's reprieve

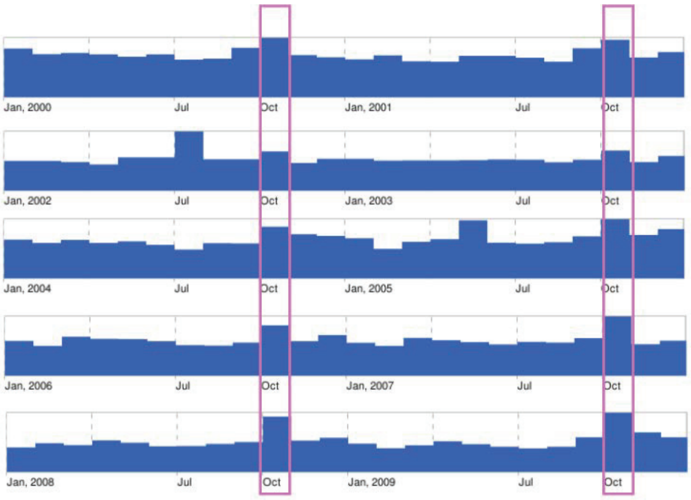
In my living nightmare, I see no light;
Neither the sliver of a crescent nor a star
Just the deep emptiness of an endless night

- The Author

# FOREWORD

I came across the now obsolete tool called Google timeline by pure serendipity while doing my literature research on breast cancer online. It plots the quantity of Google results related to breast cancer added to the cyberspace over time (Figure 0-1), and the month of October stood out like a suburban skyscraper. It then dawned on me that this burst of frenzied byte traffic must be due to the Pink Ribbon Campaign.

**Figure 0-1 Volume of Google results related to breast cancer added to the cyberspace**



Since the Breast Cancer Awareness Month was conceived in 1985, many women have for one reason or another - guilt-tripped, tempted, or otherwise - been amassing pink products for 30 days in a year. I paid premium for everything from my compact camera, printer, wetsuit, dive computer, to kitchen towel and toilet paper, which needless to say, came in different shades of pink, and blamed it all on my research project, which deals with the genetics of breast cancer. Surely, I must support the very cause I am working for?

Slowly people are starting to realize that much hype has been focused on looking for a cure, and too little attention being spent on preventing or early detection of the disease and understanding what causes cancer in the first place. In this thesis, I look into the book of life itself, scrutinizing at the DNA that defines us, for genetic differences that spell who is likely to get breast cancer, and who is not. The aim is to discover novel susceptibility markers and mechanisms, which are bits and pieces of clues essential to solving the puzzle of the disease. Knowing what makes the cancer bomb tick will ultimately be helpful in stratifying the population according to the likelihood of getting the disease, so that resources can be reallocated to screen individuals at high risk more often than those with below average risk of getting breast cancer.

October or not, the fight against breast cancer goes on.

# ABSTRACT

The main purpose of this thesis was to identify genetic risk factors using both hypothesis-based and hypothesis-free approaches.

In an attempt to identify common disease susceptibility alleles for breast cancer, we started off with a hypothesis-free approach, and performed a combined analysis of three genome-wide association studies (GWAS), involving 2,702 women of European ancestry with invasive breast cancer and 5,726 controls.

As GWAS has been said to underperform for studying complex diseases such as breast cancer, we investigated to see if the variance explained by common variants could be increased by studying specific disease subtypes. Breast cancer may be characterized on the basis of whether estrogen receptors (ER) are expressed in the tumour cells. The two breast cancer tumour subtypes (ER-positive and ER-negative) are generally considered as biologically distinct diseases and have been associated with remarkably different gene expression profiles. ER status is important clinically, and is used both as a prognosticator and treatment predictor since it determines if a patient may benefit from anti-estrogen therapy. We thus performed an independent GWAS using a subset of ER-negative breast cancer cases and all of the controls from the initial genome-wide study, and, in addition, also evaluated whether the two cancer subtypes are fundamentally different on a germline level.

Besides hypothesis-free GWAS, we also conducted hypothesis-based analyses based on candidate pathways to identify common variants associated with breast cancer. Several studies have examined the effect of genetic variants in genes involved in the estrogen metabolic pathway on mammographic density, but the number of loci studied and the sample sizes evaluated have been small and pathways have not been evaluated comprehensively. We evaluated a total of 239 SNPs in 34 genes in the estrogen metabolic pathway in 1,731 Swedish women who participated in a breast cancer case-control study.

Slightly venturing outside the genetic scope of this thesis, we looked at a breast cancer risk factor - body size - that is associated with very different postmenopausal breast cancer risks at different time points in a woman's lifetime, namely, birth, childhood, and postmenopausal adult.

The significance of these studies will be apparent when, using the new genetic and epidemiological knowledge found, we are able to classify women according to high or low risk of breast cancer on the basis of genetic disposition or other breast cancer risk factors, so that appropriate interventions and disease management decisions may be made, to ultimately reduce incidence and mortality of breast cancer.

Keywords: Breast Neoplasms, Genetic Epidemiology, Genetic Susceptibility, Genetic Predisposition to Disease/genetics*, Case-Control Studies, Genetic Association Studies, Candidate Gene Analysis, Gene Discovery, Single Nucleotide Polymorphism, Risk Factors,, Estrogen Receptors, Mammography, Body Size

## LIST OF PUBLICATIONS

I. A combined analysis of genome-wide association studies in breast cancer.
Li J, Humphreys K, Heikkinen T, Aittomäki K, Blomqvist C, Pharoah PD, Dunning AM, Ahmed S, Hooning MJ, Martens JW, van den Ouweland AM, Alfredsson L, Palotie A, Peltonen-Palotie L, Irwanto A, Low HQ, Teoh GH, Thalamuthu A, Easton DF, Nevanlinna H, Liu J, Czene K, Hall P.
*Breast Cancer Res Treat*. 2010 Sep 26.

II. A genome-wide association scan on estrogen receptor -negative breast cancer.
Li J, Humphreys K, Darabi H, Rosin G, Hannelius U, Heikkinen T, Aittomaki K, Blomqvist C, Pharoah PD, Dunning AM, Ahmed S, Hooning MJ, Hollestelle A, Oldenburg RA, Alfredsson L, Palotie A, Peltonen-Palotie L, Irwanto A, Low HQ, Teoh GH, Thalamuthu A, Kere J, D'Amato M, Easton DF, Nevanlinna H, Liu J, Czene K, Hall P.
*Breast Cancer Res*. 2010 Nov 9;12(6):R93.

III. Genetic variation in the estrogen metabolic pathway and mammographic density as an intermediate phenotype of breast cancer.
Li J, Eriksson L, Humphreys K, Czene K, Liu J, Tamimi R, Lindstrom S, Hunter DJ, Vachon C, Couch F, Christopher S, Lagiou P, Hall P.
*Breast Cancer Res*. 2010 Mar 9;12(2):R19.

IV. Effects of childhood body size on breast cancer tumour characteristics.
Li J, Humphreys K, Eriksson L, Czene K, Liu J, Hall P.
*Breast Cancer Res*. 2010 Apr 15;12(2):R23.

# CONTENTS

# TABLE OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ADHD | Attention deficit hyperactive disorder |
| AML | Admixture maximum likelihood |
| AUC | Area under curve |
| CAHRES | Cancer Hormone Replacement Epidemiology in Sweden |
| CGEMS | Cancer Genetic Markers of Susceptibility |
| CNV | Copy number variation |
| COGS | Collaborative Oncological Gene-Enivronment Study |
| CT | Cycle threshold |
| DNA | Deoxyribonucleic acid |
| EIRA | Epidemiological Investigation of Rheumatoid Arthritis |
| ER | Estrogen receptor |
| FGC | Finnish Genome Center |
| GWA/GWAS | Genome-wide association study |
| HUBC | Helsinki University Breast Cancer Study |
| KARMA | Karolinska Mammography |
| kb | Kilobase(s) |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LD | Linkage disequilibrium |
| MALDI-TOF | Matrix-assisted laser desorption/ionization time of-flight |
| MBCS | Mayo Clinic Breast Cancer Study |
| MODE | Marker of DEnsity consortium |
| NHS | Nurses' Health Study |
| PCA | Principal component analysis |
| POLR | Proportional odds logistic regression |
| PR | Progesterone receptor |
| RBCS | Rotterdam Breast Cancer Study |
| RNA | Ribonucleic acid |
| ROC | Receiving operator characteristic |
| SCAN | SNP and CNV Annotation Database |
| SEARCH | Studies in Epidemiology and Risks of Cancer Heredity |
| SNAP | SNP Annotation and Proxy Search |
| SNP | Single nucleotide polymorphism |
| SRT | SNP Ratio Test |
| WGA | Whole genome association study |

# 1 INTRODUCTION

Breast cancer is not just a lump - it's a killer disease. One in eight women will get breast cancer in their lifetime. Statistics from GLOBOCAN estimated that 458,000 women died from breast cancer globally in 2008 (Figure 1-1) [1, 2], which is equivalent to the loss of one life to the disease nearly every minute. Approximately 1,383,000 new cases of invasive breast cancer (23% of all cancers among women) were diagnosed globally in 2008 (Figure 1-2).

In developed countries, breast cancer is the leading cause of cancer death in women between the ages of 15 and 54, and the second cause of cancer death in women 55 to 74. The bulk of the women with breast cancer (77%) are over 50. In view of the large proportion of postmenopausal breast cancer cases, the focus of studies described in this thesis are on this group of women.

Breast cancer is hereditary in nature, with both genetic and non-genetic risk factors (we inherit more than just genes from our parents; we also inherit lifestyle to a certain extent). It has been reported that 27% of breast cancer risk may be explained by heritable factors [3]. It is, however, suggested that genetics plays the larger role. In sets of twins with at least one twin with breast cancer, twin pairs have been found to be concordant for breast cancer in monozygotic pairs more than in dizygotic pairs.

Rare, high-penetrance and high-risk variants, such as *BRCA1*, *BRCA2* and *TP53*, and rare, intermediate risk variants, such as *PTEN*, *CHEK2*, *PALB2* and *BRIP1*, can only explain 27% of the *excess familial risk*[1] of breast cancer [4]. Common variants identified through recent genome-wide association studies (GWAS) have currently shown to be responsible for a further 5%, leaving more than two-thirds of genetic risk unaccounted for [4]. Despite the increased understanding of genetic predisposition to breast cancer in recent years, the field remains fertile for the discovery of novel genes/loci to better understand the architecture of breast cancer.

With the completion of the Human Genome Project and rapid technological advances, we are in a good position to scour the genetic landscape for the elusive variants that, though common, has only small effects, or variants that only exert effects in the presence of other risk factors. The aim of this work is to identify common variants that predispose to the risk of breast cancer, and increase the explained variance, using a variety of analyses and approaches.

B~~r~~ea~~s~~t Breast B~~r~~east[2]

The overarching goal is to one day be able to classify women according to high or low risk of breast cancer on the basis of genetic disposition or other breast cancer risk factors, so that appropriate interventions and disease management decisions may be made, to ultimately reduce incidence and mortality of breast cancer.

[1] The increased risk of developing the disease in a relative of an affected individual.

[2] Anagrams for the word breast - Beat Breast Beast

**Figure 1-1 Global breast cancer mortality in 2008**
Coloured bar indicates age-standardized incidence rates per 100,000. Source: [1]

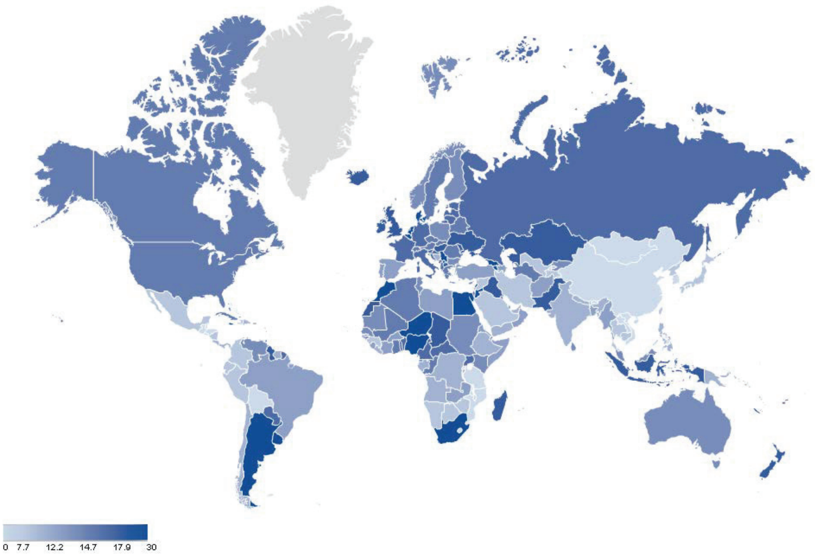
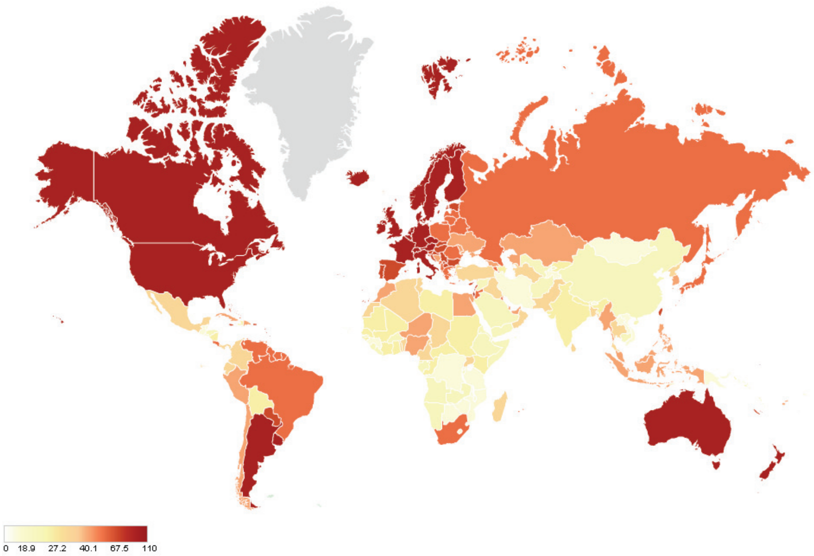


**Figure 1-2 Global breast cancer incidence worldwide in 2008**
Coloured bar indicates age-standardized incidence rates per 100,000. Source: [1]



# 2 BACKGROUND

## 2.1 BREAST CANCER STATISTICS

Breast cancer is the most common cancer among women (Figure 2-1) [1]. More than one million women are diagnosed with breast cancer globally every year [2]. Between 8 and 12 percent of women in the western world will be diagnosed with the disease during their lifetime and the incidence is increasing [2, 5]. Breast cancer risk increases with age. The incidence of breast cancer increases with age and doubles every 10 years until the menopause when the rate of increase slows (Figure 2-2). Approximately 25% of breast cancer cases affect women under the age of 50, 50% occur in women between ages of 50 and 69, and the remaining develop in women 70 years and older.

**Figure 2-1 Most common cancers in women**
Source: [1]

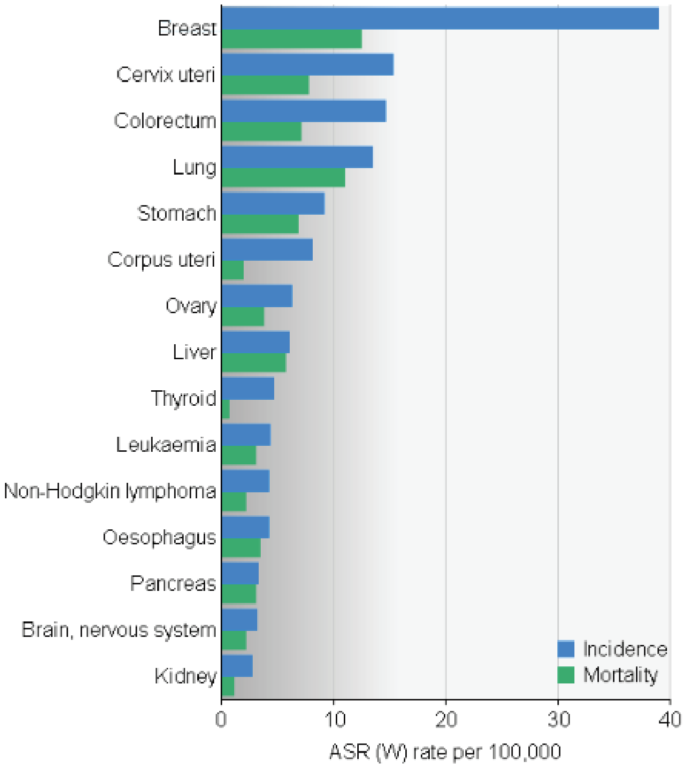
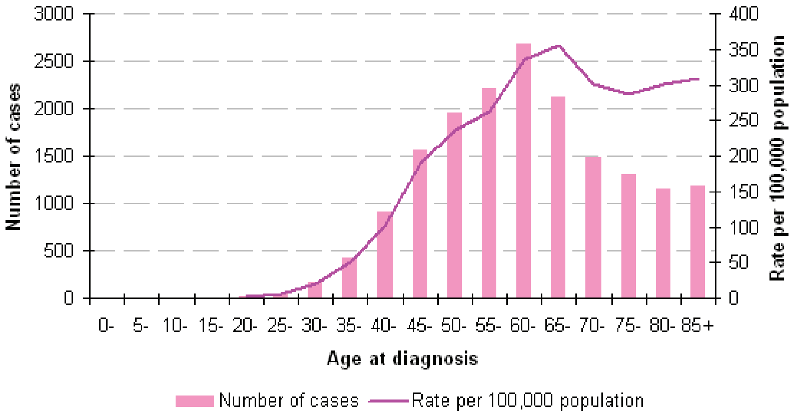


**Figure 2-2 Number of new breast cancer cases, Nordic countries, 2007**
Source: (4)



## 2.2 GENETICS OF BREAST CANCER

The main causative culprit behind sporadic cancers is the environment. The etiological make-up of a heterogeneous and complex disease such as breast cancer is diverse [6], and includes age, geographical location, lifestyle factors, environmental factors, and hormonal factors, among others [6, 7].

Genetics is also known to play a part. Although all cancers are familial[3] to a certain degree, inherited genetic factors have been reported to only make a minor contribution to the susceptibility of most types of site-specific cancers [8]. However, the heritable component of breast cancer derived from twin studies is estimated to be relatively high (~27%) [3], and genetic effects have been calculated to explain almost 30% of the total variability of propensity to breast cancer [9], making the disease a good candidate for gene hunts.

### 2.2.1 SNP-ing the genome

The DNA alphabet consists of four letters or nucleotides, A, T, G or C. Single nucleotide polymorphisms, or SNPs (pronounced "snips"), are single letter alterations in the deoxyribonucleic acid (DNA) sequence. For example a SNP might change the DNA sequence **T**AGCAT to **G**AGCAT. A variation at a single position is considered a SNP if it occurs in at least 1% of the population, and is thus sometimes referred to as a "common variant".

SNPs make up the bulk of all human genetic variation (~90%), and are densely distributed across the 3-billion-base human genome, occurring every 100 to 300 bases. Most SNPs (every two out of three) involve the replacement of cytosine (C) with thymine (T). The repercussions of having a variant SNP can vary, as SNPs

[3] Familial risk of a disease is a measure of its clustering in family members. Commonly, familial risk is defined between those who have a relative (e.g., parent or sibling) with cancer compared to those whose relatives are free from cancer, given as a familial relative risk or familial standardized incidence ratio (SIR)

can occur in both the exonic (gene coding) or intronic (non gene coding) regions of the genome. The vast majority of SNPs have no direct contribution to a change in disease status, but because a SNP may be linked to another functional SNP by means of shared underlying genetic architecture, they are often studied as markers that could help determine the likelihood that someone will develop a particular disease or trait.

SNPs are not only useful in identifying meaningful disease-related hotspots by being guilty by association. Another interesting concept that involves the ability of SNPs to determine disease outcome is host genetics. In a set of breeding studies performed on mice, it was found that the same cancer-causing stimulus in the male mice - the expression of the polyoma middle-T antigen transgene – manifested different capacities to form tumours in their offspring, when mated to female mice of varying genetic backgrounds [10]. Collectively, the genetic background defined by SNPs may be important in modifying the effects of other genetic and non-genetic breast cancer risk factors via interactions.

Because SNPs are so plentiful, a large number of such variants are usually studied simultaneously. In genetic epidemiology, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study), is an examination of all or most of the genes (the genome) of different individuals to see how much the genes vary from individual to individual. Different variations, such as SNPs, are then associated with different traits, such as diseases. In humans, this technique has found associations of particular genes with diseases such as age-related macular degeneration [11], diabetes [12], and leprosy [13], among many others. Due to the rapid increase in the number of GWAS, online resources exist to curate and index the SNP-trait associations extracted from published literature [14].

### 2.2.2 Breast cancer susceptibility loci identified through GWAS

To date, there are ~27 instances of SNPs identified as “breast cancer susceptibility loci”. The list in Table 2-1 does not comprise of unique SNPs. The same SNPs, such as rs2981582 and rs3803662, may have been identified independently in different GWAS, and thus appearing multiple times. In addition, the associations might not be wholly independent. In population genetics, linkage disequilibrium (LD) is “the nonrandom association between two or more alleles such that certain combinations of alleles are more likely to occur together on a chromosome than other combinations of alleles” (The American Heritage® Medical Dictionary). For example, rs1219648 and rs2981582 located in the FGFR2 gene are in perfect LD (r2 = 1) [4], and are thus perfect surrogates for each other.

[4] r2 is a measure of linkage disequilibrium which ranges between 0 (when they are in perfect equilibrium) and 1 (when the two markers provide identical information). It is sometimes used to measure a loss in efficiency when marker A is replaced with marker B in an association study.

**Table 2-1 List of common breast cancer susceptibility SNPs and the corresponding genes they are associated with.**

| PMID | SNP | CHR | BP | Alleles | GENE |
|---|---|---|---|---|---|
| 19330030 | rs11249433 | 1 | 120982136 | C/T | INTERGENIC |
| 17529974 | rs13387042 | 2 | 217614077 | A/G | INTERGENIC |
| 19330027 | rs4973768 | 3 | 27391017 | C/T | *SLC4A7* |
| 18438407 | rs10941679 | 5 | 44742255 | A/G | INTERGENIC |
| 18438407 | rs4415084 | 5 | 44698272 | C/T | INTERGENIC |
| 17529967 | rs889312 | 5 | 56067641 | A/C | INTERGENIC |
| 19219042 | rs2046210 | 6 | 151990059 | C/T | INTERGENIC |
| 20453838 | rs3757318 | 6 | 151955806 | A/G | *C6orf97* |
| 17529967 | rs13281615 | 8 | 128424800 | A/G | INTERGENIC |
| 20453838 | rs1562430 | 8 | 128457034 | A/G | INTERGENIC |
| 20453838 | rs1011970 | 9 | 22052134 | G/T | INTERGENIC |
| 20453838 | rs10995190 | 10 | 63948688 | A/G | *ZNF365* |
| 17529973 | rs1219648 | 10 | 123336180 | A/G | *FGFR2* |
| 20453838 | rs2380205 | 10 | 5926740 | C/T | INTERGENIC |
| 17529967 | rs2981582 | 10 | 123342307 | C/T | *FGFR2* |
| 19536173 | rs2981582 | 10 | 123342307 | C/T | *FGFR2* |
| 19536173 | rs3135718 | 10 | 123343859 | A/G | *FGFR2* |
| 20453838 | rs704010 | 10 | 80511154 | A/G | INTERGENIC |
| 19536173 | rs7895676 | 10 | 123323987 | C/T | *FGFR2* |
| 17529967 | rs3817198 | 11 | 1865582 | C/T | *LSP1* |
| 20453838 | rs614367 | 11 | 69037945 | C/T | INTERGENIC |
| 20453838 | rs909116 | 11 | 1898522 | C/T | *TNNT3* |
| 19330030 | rs999737 | 14 | 68104435 | C/T | *RAD51L1* |
| 17529967 | rs12443621 | 16 | 51105538 | A/G | *TOX3* |
| 17529967 | rs3803662 | 16 | 51143842 | C/T | *LOC643714* |
| 17529974 | rs3803662 | 16 | 51143842 | C/T | *LOC643714* |
| 17529967 | rs8051542 | 16 | 51091668 | C/T | *TOX3* |

One of the genes associated with breast cancer, fibroblast growth factor receptor 2 or *FGFR2*, is a good example of a GWAS-identified locus that has been implicated in the disease development of breast cancer [15, 16]. The association signals from the highly significant hits of the GWAS brought attention to a specific region on chromosome 10, which previously have not been linked to breast cancer.

Through *fine-scale genetic mapping*[5] of the region, it has been possible to narrow the causative locus to a haplotype of eight strongly linked SNPs spanning a region of 7.5 kilobases (kb) in the second intron of the *FGFR2* gene, and more studies are underway to identify the true causative variant [17].

### 2.2.3 Prediction is very difficult, especially if it's about the future.

- Niels Bohr

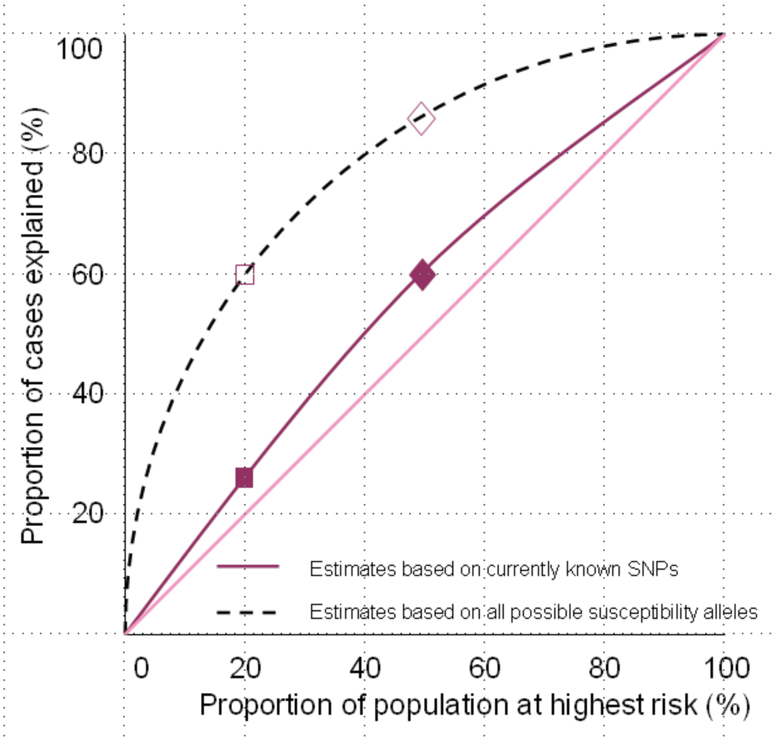
There have been attempts at understanding how useful the SNPs mentioned above could be for predicting breast cancer risk and aid in the target prevention of breast cancer [18].

[5] Fine-mapping involves the identification of markers that are very tightly linked to a targeted gene.

To gauge whether a predictive model is performing well, we can plot a receiver operating characteristic (ROC) curve. The area under the curve (AUC) measures discrimination, that is, the ability of the test to correctly classify those with and without the disease. An area of 1 represents a perfect prediction; an area of 0.5 is not informative at all, that is, the results of the prediction model are no better than randomly flipping a coin.

Figure 2-3 is an example of a ROC curve and shows the difference in predictive power achieved by using seven of the currently known SNPs [18] (denoted by a thick red line). The black dashed line shows the theoretical scenario when all possible susceptibility alleles are included in the model. The pink null line illustrates a scenario where the SNPs have negligible value in explaining the proportion of breast cancer cases in a population.

**Figure 2-3 Proportion of cases of breast cancer explained by the proportion of the population at highest risk for breast cancer.**
Source: [18]



Assume that there are 100 breast cancer cases in a population of 1000 women. If we were to genotype the entire population for the seven breast cancer susceptibility loci used in the example of Pharoah et al. [18], and rank them according to their genetic risk profiles, we would expect to identify a quarter of all breast cancer cases (25/100) amongst the 200 women with the highest risk as determined by the seven susceptibility loci (solid square). Similarly, among 500 women with the highest genetic risk scores, we would expect to find 60% of all the breast cancer cases (60/100 women, solid diamond).

In an ideal world where we have full knowledge of all the variants that predispose a woman to breast cancer, and if we genotyped the entire population, we could expect to find more than half of all the breast cancer cases (60/100) among the women with the highest genetic risk profiles (20% of all women, unfilled square). That is a huge jump from the ~25% explained using only seven of the currently known SNPs.

Similarly, with knowledge of all breast cancer variants, we would expect to find more than 80% of all breast cancer cases (80/100) among half of the women with the highest risk profiles (unfilled diamond). That is, if we knew the genetic risk profiles of the entire population, we can selectively apply prevention measures to only half the women (i.e. screen the women at high risk more frequently, or provide chemoprevention therapy), yet prevent more than 80% of all breast cancer cases. Besides being easier on national health budgets, fewer women would need to experience unnecessary hassle or undesirable side effects of chemoprevention, for example.

Despite the efforts of several independent GWAS, little progress has been made from the solid line to the dotted line in Figure 2-3. In a recent large prospective study consisting of 10,306 women with breast cancer and 10,393 women without breast cancer, the effects of 14 breast cancer susceptibility loci identified through the various GWAS efforts have been estimated [19]. It was found that women who had the highest risk scores (highest quintile) were twice as likely as those who had the lowest risk scores (lowest quintile) to get breast cancer.

Although the results were encouraging, the genetic risk score was not much better than family history in predicting breast cancer risk. Wacholder et al. [20] found that traditional breast cancer risk factors (i.e. age at menarche, age at first live birth, number of previous biopsies, and number of first-degree relatives with breast cancer, which are considered in the Gail Model [21]), showed an AUC of 58.0%. The inclusion of the newly discovered genetic factors only modestly improved the performance of risk models for breast cancer, increasing the AUC to 61.8%. If the improvement was only better by 3.8 percentage points, why should we even consider genotyping the entire population, when we can simply ask women to fill in answers to a few questions online and providing them with an instant feedback of their breast cancer risk?

At present, it is unlikely that such polygenic risk scores would be used in population-based screening programs. However, as more SNPs are identified, the predictive value of these markers will clearly improve, and may prove to be useful in understanding biological mechanisms behind breast cancer etiology.

## 2.3 **MISSING HERITABILITY**

It seems a bit strange that more predictive power can't be squeezed out of the nine independent GWAS performed. If genetics really play a large part in the heritability of breast cancer, then maybe we are not looking hard enough. Below, I summarize some of the possible explanations for this "missing heritability".

**Rare variants**. GWAS has its limitations. Rare associations are typically missed by current GWAS methods [22]. While common variants identified through recent GWAS to date can explain only ~5% of the familial risk of breast cancer, the known rare, high-penetrance breast cancer variants with large effects, such as *BRCA1*, *BRCA2* and *TP53*, and rare, intermediate risk variants, such as *PTEN*, *CHEK2*, *PALB2* and *BRIP1*, account for ~27% [4]. A large bulk of the genetic landscape of breast cancer remains unmapped, and the reason behind this missing heritability has been much discussed, debated and deliberated [23-25]. Rare variants that are yet identified, which occur in between one to five percent of the population, with large effect sizes are among the many proposed candidates for explaining this missing heritability.

However, from the latest developments, we are seeing compelling evidence that rare variants do NOT explain disease variance over and above that of common variants. For example, Momozawa et al. [26] identified low frequency coding variants through resequencing of positional candidates conferring protection against inflammatory bowel disease in *IL23R*, but concluded that rare coding variants in positional candidates do not make a large contribution to inherited predisposition to Crohn's disease. Rare variants, if I may say so, appear to be a fashion trend; they come and go like a bad case of the flu.

**Genetic mutations do not usually act alone**, and conditions attributed to a single genetically dominant and almost fully penetrant variant, such as Huntington's disease, are rare. Since breast cancer is a complex disease, it does not obey the single-gene dominant or single-gene recessive Mendelian law. Rather, genes tend to work in groups, a phenomenon known as gene-gene interaction or epistasis. A small change in a gene may modify the effects of other genes. By looking at only single marker effects, effects due to such interactions of genes are not accounted for.

**Genetic heterogeneity** is the phenomenon that a single phenotype or genetic disorder may be caused by any one of a multiple number of alleles or non-allelic (locus) mutations [27]. By performing a combined GWAS, or a meta-analysis of independent GWAS, and looking at the combined p-values of single markers, we may miss out association signals which are important within individual populations.

**Interactions are not limited to between genes and genes only**. On top of the need to consider the effect of genes in the presence of other genes, one needs to also factor in environmental influence (gene-environment interaction or G × E). A classic example is a human genetic condition known as phenylketonuria, which is caused by mutations to a gene coding for a particular liver enzyme [28]. Left untreated, a defect in the metabolism of a specific protein building block known as phenylalanine causes severe mental retardation, epilepsy and behavioral problems. By changing the environmental exposure, or in this case, restricting phenylalanine in diets for newborns screened positive for this condition, most affected infants grow up leading normal lives.

**Increase resolution.** Another strategy for uncovering hidden heritability is to examine DNA in more detail. We are currently speed-reading the book of life at best, picking out only words we deem to be relevant to our understanding of the genome. For instance, to maximize the investment in genotyping and statistical power, a subset of informative SNPs selected based on linkage disequilibrium (also known as “tag SNPs”) is often used in GWAS [29, 30]. Although it is possible for one to extract and comprehend the main storyline of a novel by reading only one in every ten words, certain savoury details may still be missed. Small informative footnotes that might also be easily missed include mitochondrial DNA. Unless denser microarrays or whole genome sequencing technologies are applied, we might never tease out information hidden away in the rest of the genome.

**Increase statistical power.** Besides looking at more variants, we also need to look at more people. Statistical power to detect an effect is limited by the sample size, or the number of individuals included in the study. For example, height is a complex trait that is possibly determined by hundreds of loci with very modest effect sizes which are difficult to detect without sufficient statistical power. More than a hundred thousand DNA samples were analyzed in recent GWAS efforts by the GIANT Consortium to identify loci associated with body mass index [31] and height [32].

**Structural variation**. Although SNPs are the most predominant form of genetic variation, they are not the only form. Besides single-letter differences, two individuals may also be different on the structural level of DNA – deletions or duplications of DNA regions, inversions etc. Copy number variations, or CNVs, are similar to repetitions or deletions of blocks of text in the story. Jane’s genetic instructions could read “I am very pretty”, while Mary’s could be “I am very, very pretty”. The extra copy of “very” in the text would mean that Mary is probably prettier than Jane, because it is coded so. Another glitch which may occur is when an extra copy is present in the wrong place – “I am very prverytty” would confuse the system and no prettiness would be coded as a result. An example of CNV in humans is the starch-digesting enzyme amylase. Populations which consumed starchy diets (European Americans, the Japanese, and Hadza hunter-gathers) were found to have more copies of the gene than populations which kept to a low-starch diet (two rainforest hunter-gatherers, the Mbuti and Biaka and two pastoralist groups, the Datog and Yakut) [33].

**Non-genetic changes.** The actual impact of a gene on the end phenotype is also subjected to non-genetic changes, such as epigenetic and post-translational modifications of gene expression. Processes such as histone acetylation and deacetylation function as a switch between repressive and permissive chromatin to govern transcriptional activity [34]. Other epigenetic processes such as DNA methylation and histone modification are associated with gene-silencing-associated events [35]. In addition, small non-coding ribonucleic acid (RNA), called microRNA, can post-transcriptionally modulate the expression of more than a third of the coding messenger RNAs without changing the underlying genetic code.

**Restrictive assumptions of heritability estimates**. Heritability estimates are exactly what they are – *estimates*. In a commentary by Rose [36], several misconceptions over the definition of the term were discussed. The measure refers to the proportion of phenotypic variation attributable to all genetic causes in a population within a population in a specific environment; if the environment changes, the heritability measure changes. In addition, the measure cannot be used to explain causes of differences between populations. Implicit in the heritability measure is the assumption that the contributions of genes and environment are additive, but it is also possible that interactions occur on a multiplicative scale. The successful application of heritability estimates outside the narrow range of circumstances for which it was originally derived is thus limited.

**"Garbage in; garbage out."** The quality of the data that is being scrutinized is of utmost importance. In order to pool samples together in gigantic consortia to achieve statistical power, a trade-off is often made with phenotypic precision. Although measurement error rate is low for genetic polymorphisms, the same cannot always be said for the outcomes of interest. As disease definitions are typically not clear cut, the definition of what constitutes a “case” in collaborative GWAS is at times arbitrary, especially for spectrum disorders, such as autism, attention deficit hyperactive disorder (ADHD) and schizophrenia [37]. Rigorous, adequately powered studies homing in on well-defined subtypes of heterogeneous diseases such as Parkinson’s disease [38] or breast cancer [39] may be required to identify genetic variants associated with the different subtypes, which could be etiologically distinct.

This thesis explores the problem of ambiguous phenotypes obscuring GWAS results in more detail. Breast cancer may be characterized on the basis of whether estrogen receptors (ER) are expressed in the tumour cells (described in more detail in the following section). ER status is important clinically, and is used both as a prognostic indicator and treatment predictor since it determines if a patient may benefit from anti-estrogen therapy. Approximately one third of all breast cancers are ER-negative, and cancers of this ER subtype are highly age-dependent and generally have a more aggressive clinical course than hormone receptor-positive disease.

## 2.4 ORIGINS OF ER-NEGATIVE BREAST CANCER

Estrogens act on target tissues by binding to parts of cells called estrogen receptors (ER) which normally reside in the cell’s nucleus, along with DNA molecules [40]. In the presence of estrogen, ER triggers gene activation to induce changes in cell behaviour. In some target tissues, estrogen plays an important role in causing cells to grow and divide, a process called cell proliferation. Although this ability to stimulate cell proliferation is one of estrogen’s normal roles, it can also increase a woman’s chance of developing a cancer in the target tissue where ER is expressed. Estrogen receptors are not always expressed in cancer cells arising in the breast; those breast cancers that do have ER are said to be “ER-positive,” while those breast cancers that do not possess ER are “ER-negative.”

Overall, the evidence appears overwhelming that ER-negative breast cancers originate from ER-positive precursors [41]. Allred et al. [41] summarized evidence supporting the opinion that ER status can switch from one subtype to another, in either direction, from epidemiological, histological/pathological and molecular aspects. Firstly, increased exposure to estrogen has been associated with increased breast cancer risk for both ER-positive and ER-negative breast cancers. In addition, a decrease in estrogen exposure in *BRCA1* mutation carriers is correlated with a decreased risk of breast cancer, also for both ER-positive and ER-negative breast cancers. Secondly, early stage breast cancers tend to be predominantly ER-positive, with progressively more ER-negative tumours among women with late-stage cancers. ER-positive precursors have also been detected in ER-negative tumours in the same patient. Lastly, molecular mechanisms such as *MAPK* activation or hypermethylation of ER promoters have been shown to experimentally alter ER status in a reversible manner.

There has been considerable debate as to whether breast cancers of different ER subtypes really share a common root (i.e. ER-positive precursors). Allred et al. [41] presented arguments for this alternative view, which is now generally regarded as the mainstream view. For example, anti-estrogens, such as tamoxifen, which blocks ER in breast tissue, are only effective as chemoprevention therapy against ER-positive cancers. In a series of seminal articles, breast cancer was found to consistently show several distinct gene expression patterns, each of which was coined a "molecular portrait of cancer", or breast cancer subtype [42-44]. ER status was one of the key determining factors of this classification.

## 2.5 MAMMOGRAPHIC SCREENING

### 2.5.1 A specific kind of X-ray

A mammogram is a special X-ray examination of the breast. The first sign of breast cancer usually shows up on a woman's mammogram before it can be felt or any other symptoms are present. Early detection of breast cancer through yearly mammography, together with monthly breast self-examination, offers the best chance for survival. Over 96% of women who find and treat breast cancer early (Stage 0/I, or when cancer is confined to the breast [45] have an excellent chance of complete recovery and of remaining cancer-free after five years. Otherwise, the five-year survival after diagnosis is 89% for all breast cancers [45]. As a result of the excellent chance of complete recovery, more than 1.7 million women who have had breast cancer are still alive in the United States.

### 2.5.2 Limitations of mammography

Early cancer detection, however, comes with a price. Mammography is simply too good at finding irregularities in the breast. The bumper crop of breast cancer cases among women between 40-65 years of age, which coincides with the window for mammography screening (Figure 2-2). This increase in cases could be attributed to the detection of latent breast cancers. The question is then whether all cancers demand equal attention. Do small, early-detected, non-invasive in situ carcinomas signal big problems to a woman's health?

The magnitude of overdiagnosis from randomized trials ranges from 10 to 52% [46-48]. Although the estimates differ substantially among studies, the evidence for overdiagnosis of breast cancer with mammography screening is consistent and strong. It is a source of grave concern that many women are being told the devastating news that they have a cancer, or being treated with unnecessary therapy that is often fraught with serious side effects, when in fact there is considerable chance that a mammographic abnormality, when left untreated, may never advance into a deadly malignant tumour.

There are also times when mammography does not deliver. Mammographic screening sensitivity is affected by the amount of dense tissue present in the breast [49]. Against a background of dense tissue, abnormalities such as tumours may be “masked”, making them harder to detect. Since a woman’s breasts decrease in density with age, mammography is an ideal technique for screening for abnormalities in breasts of older women. It has also been recommended that women in high risk groups with dense tissue patterns should go for more frequent screens and/or with more views per breast, or be prescribed chemoprevention [49] to avoid missing suspicious radiographic lumps.

### 2.5.3 Mammographic density is a measure of risk

Limitations aside, in addition to its diagnostic virtues, the proportion of radiographically dense (white areas) to non-dense, predominantly fatty, tissue (dark areas), on a mammogram is an independent risk factor and one of the strongest indicators of breast cancer risk [50]. Several studies have shown that women with extensive dense tissue are at between four to six times higher risk of developing the disease than women of similar age with lower mammographic density [51, 52]. Examples of other risk factors found to be indisputably linked to certain diseases are smoking to lung cancer, and recurrent reflux to esophageal cancer. To put things in perspective, the odds ratio for lung cancer in current United States smokers relative to nonsmokers was 40.4 [95% confidence interval = 21.8-79.6] [53], and recurrent symptoms of reflux are associated with a 7.7-fold [95% confidence interval = 5.3-11.4] increase in risk of getting esophageal cancer [54]. On the other hand, having a first degree relative with a history of breast cancer only increases one’s risk of getting breast cancer by approximately two-fold [55].

There are various measures of mammographic density. Wolfe was the first to introduce the first qualitative classification of breast tissue patterns in 1976 [56]. The four classification categories - N1, P1, P2 and DY – describe a breast that is almost entirely fat, a breast with scattered fibroglandular densities, a heterogeneously dense breast, and an extremely dense breast, respectively [57]. Tabár et al. [58] later proposed a modification to Wolfe’s classification by separating Wolfe’s N1 pattern into two subgroups. Wolfe also quantified on a continuous scale the percentage of radiologically dense areas on a mammogram with the use of a polar planimeter [59]. This method was later modified into the BI-RADS system and adopted for use in clinical radiology practice in the USA [49]. Several semi-automatic computer-assisted techniques are also available to assess mammographic density quantitatively [60, 61]. Computer-aided thresholding programs, such as Cumulus, are currently seen as the accepted standard for measurement of mammographic density.

Overall, there is substantial agreement across different assessment methods in determining high-risk (high density) versus low-risk (low density) mammographic patterns [49, 62]. Measurements by quantitative scales, such as Boyd and BI-RADS, are highly reproducible, with almost perfect agreement. On the other hand, methods which rely on ratings of parenchymal tissue patterns by an observer, i.e., Tabar and Wolfe, perform well, but have only good agreement.

Besides mammography, other techniques used to capture abnormalities in the breast include ultrasound tomography [63] and magnetic resonance imaging [64]. It has been proposed that such alternative methods of imaging may complement the characterization of breast density by mammography to improve breast cancer risk prediction and disease prevention [65, 66]. However, due to the dual considerations of cost and ease of measurement [67], mammography is the most prevalent technique used for the characterization of breast density.

### 2.5.4 Genetics of mammographic density

Twin studies have estimated the heritability of the mammographic density trait to be between 60-67% [68]. Evidence for a genetic influence also comes from other studies on family history, familial aggregation and segregation analyses. As underlying risk factors of complex diseases are likely to share genetic variants with the disease itself [69], unravelling the genetics of mammographic breast density may offer insights into the carcinogenesis of breast cancer. As a phenotypic manifestation such as mammographic density is more proximal to the endpoint (i.e. breast cancer) on the causal chain than genetic polymorphisms, the examination of this trait is likely to narrow down the possible genetic and environmental factors influencing the disease outcome. Hence, attempts to identify genetic determinants of mammographic density may be a more focused approach, both more powerful and more efficient, for studying the etiology of breast cancer.

Perhaps against expectations, attempts at finding a genetic link between known common susceptibility loci of breast cancer (from GWAS) and mammographic density have mostly been inconclusive [70-72]. However, a recent Australian study revealed a positive connection between the same breast cancer SNPs and mammographic density [73]. Although the scientific media immediately homed in to celebrate this “expected” finding [74-76], in view of the past endeavours, the results should be interpreted with caution. Nevertheless, a meta-analysis of five genome-wide association studies of percent mammographic density and reported an association with rs10995190 in *ZNF365* (combined $P=9.63\times10^{-10}$) (manuscript accepted for publication in Nature Genetics). The authors claimed that this finding may partly explain the underlying biology of the recently discovered association between common variants in *ZNF365* and breast cancer risk [77].

Besides breast cancer SNPs identified using GWAS, mammographic density has also been studied in relation to genetic variation in pathways associated with breast cancer, such as steroid hormone [78-80], insulin-like growth factor (*IGF*) [81-83] and vitamin D pathways [84]. Genetic polymorphisms related to estrogen metabolism are of special interest, as a woman’s mammographic density profile correlates closely with hormonal exposure. A woman goes through menopause when her ovaries naturally stop producing estrogen and cease to function. Mammographic density has been shown to be inversely associated with age, with the largest declines observed between the years of menopause [85]. Certain regimens of hormone replacement therapy taken to counter menopausal symptoms have also been found to buffer the drop in mammographic density [86, 87].

Knowing one’s genetic predisposition to breast cancer enables a woman at a moderately increased or high risk to be active in secondary prevention of the disease (start screening at a younger age, schedule screenings more often, counselling etc). Screening women with higher than average breast cancer risk more often than women with below average breast cancer risk would also be more cost-effective for public health sectors [88, 89].

## 2.6 EPIGENETICS

### 2.6.1 Reading deeper into the book of life

The English language can be quite peculiar sometimes. Why is “argue” not pronounced as “arg” when “vogue” is pronounced as “vog”? Even more puzzling is the broad range of meanings that some words can possess, depending on the situations they are used in, where they are used (geographically), or where they are positioned in a sentence. For example, you might get shortchanged at the pump if you thought a gallon of petrol in the United Kingdom (4.54609 L) is equivalent to a gallon of gas in the United States (3.78541 L). “Boot” can be used as a verb to mean starting up a computer, or it could mean something on your foot or a car.

For the same reason, the human genome (Human Genome Project, 2003) [90] is neither just an alphabet book that came with a hefty price tag of nearly $3 billion (USD), nor should it be taken only at face value.

The term to describe changes in gene activities which do not involve alterations to the genetic code is “epigenetics” [6]. Traditionally, genetic variation has always been pinned as a culprit behind everything from a difference in eye color or height to a marker for a dreaded disease. However, the fact that every cell in our body shares the exact same genetic code, yet a cell from the surface of the skin can look rather different from a cell swapped from our tongue, is a strong hint that something else shapes development besides changes of the A-T-G-C kind. The same mechanism that acts above the DNA level to affect gene expression (and hence the prefix *epi-*) also explain why identical twins, who are virtually genetic Xerox copies of each other, may not always be respond in the same way under the same conditions (e.g one may develop cancer, the other may not) [91].

If DNA does not spell out one’s destiny, we ought to look beyond the genetic code. Depending on the ambient environment, epigenetics at work means that good genes can be silenced and bad ones jump-started, and vice-versa, and the effects of such changes can linger around for different lengths of time. The effects could be transient, like how short-term memories are formed and erased in our brains [92], or it could be life-changing, like some peculiar non-genetic sex determination systems that act in accordance with various environmental cues. For instance, many fish species such as clownfish or wrasses switch sex over the course of their lifespan depending on the social structure within their fish clans [93]. The epigenetic mechanisms underlying development or modification of reproductive systems are due to 1) changes in protein or mRNA concentration and targeting; 2) modification of protein trafficking and/or retention, or 3) post-translational modifications [94].

[6] The study of heritable changes in gene function that do not involve changes in DNA sequence.

### 2.6.2 It is often heard that a butterfly flapping its wings in South America can affect the weather in Central Park.

Very often, epigenetic marks are limited to a single generation of an organism [95]. Widespread epigenetic erasure occurs when gametes[7] are formed during meiosis[8]. Memories get reset to a blank slate when a baby gets born, and newly hatched clown fishes start off as males (or females, depending on which species), until appropriate environmental cues present themselves again. However, experiments in non-primate models have produced striking results on non-genetic inheritance. Records show that such epigenetic effects can be maintained through 13 to 40 generations in fruit flies [96] and bacteria [97], respectively, even though the offspring were not exposed to the external stimuli. In humans, it has been documented that a single winter of binge eating as a youngster could spell an earlier death for one's grandchildren [98-100]. Perhaps the tall tales of how a giraffe got its long neck from a short one (within a generation or two) by Lamarck, often said to have been denounced by Darwin's superior theory of evolution, deserve a reprieve.

## 2.7 I SEE "U"

We are programmed to be "just nice" - behold the Swedish word *lagom*. Very often we hear the wise adage saying that "too much of something is not good for you". Yet everyone knows that too little of something can be problematic too. In scientific lingo, this non-linear relationship may be classified as either a "J-" or a "U-shaped" association. Biological examples of such associations are abundant. It has been reported that being too skinny or too fat increases one's chances of dying [101]. Moderate alcohol intake has also been suggested to be protective against heart diseases, highlighting the possible adverse effects of nutritional inadequacy and excess [102].

What is good for you now may not be good for you later. Effects of external stimuli are further obfuscated by an additional dimension – time. Most of us would have had encountered major crossroads in life where our actions would lead to serious consequences and cause lasting impact, be it choosing a college, or deciding on a career path. Biologically, we are vulnerable to critical "windows" of development as well, and some important stages of life include fetal, infant, childhood, adolescence and adult.

The damage caused by environmental insults is highest when developing organisms undergo rapid growth and differentiation [103]. The breast is especially vulnerable during periods of hormonal upheaval: fetal development, puberty, pregnancy, and postmenopause [104]. For example, data from Japanese atomic bomb survivors suggests that sensitivity to radiation is highest among children or adolescents who are nearing puberty [105, 106]. In addition, while pregnancy and childbirth decreases the risk of breast cancer in the long run, the first pregnancy

[7] A mature sexual reproductive cell, as a sperm or egg, that unites with another cell to form a new organism.

[8] The special process of cell division in sexually reproducing organisms that results in the formation of gametes, consisting of two nuclear divisions in rapid succession that in turn result in the formation of four gametocytes, each containing half the number of chromosomes that is found in somatic cells.

has been linked to a transient spike in risk [107]. This is hypothesized to be due to interplay of a detrimental effect caused by intense cell growth activity in the breast, and the eventual protective effect mediated by the terminal differentiation of stem cells [107].

### 2.7.1 The "U" in growth patterns and the risk of breast cancer in women

A non-linear relationship has also been observed between anthropometric measures of body size and breast cancer risk. There is evidence that factors influencing fetal, childhood, and adolescent growth are important independent risk factors for breast cancer in adulthood [108]. Table 2-2 shows a selection of studies investigating different anthropometric measures and risk of adult breast cancer. The effect of such measures on breast cancer risk over the course of a woman's life may be described as "J"- or "U"-shaped.

### 2.7.2 Why would that be so?

"*There are many events in the womb of time which will be delivered*" (Othello: I, iii). The life of a baby starts before it enters the world. A baby's size is pegged to the risk of getting breast cancer many years into adulthood: A big baby is predisposed, while a small baby is less predisposed [108-110]. The findings of some studies suggest that the size of a baby reflects the extent of in utero hormone exposures, and a high dose of endogenous hormones, such as estrogen, so early in life may hardwire the little one's system to be vulnerable to breast cancer in adulthood (96, 97). The actual mechanisms responsible for such predisposition remain to be elucidated.

Others have speculated that a baby's anthropometric features can mediate the number of rare somatic stem cells in a manner largely independent of estrogen [111]. Stem cells are immortal, and capable of persisting into adult life. Such long lifespans make breast stem cells to be prominent targets for carcinogenesis, and any genetic frailties harboured could impact breast cancer risk later on in adulthood. Nevertheless, it has been suggested that genetic background plays a part in modifying the positive association of birth weight with adult breast cancer [112].

"*The offices of nature, bond of childhood*" (King Lear: II, iv). Childhood body size has been consistently shown to affect future breast cancer chances. From the positive association of body size at birth with breast cancer, the relationship is inversed during childhood years and young adulthood, indicative of a protective effect [108, 113-116]. It has been reported that nutrition in early life and childhood has the potential to change chromatin structure, to modify gene expression and to modulate health in adult life [117]. Hilakivi-Clarke [118] summarised in a review several perspectives on special windows of mammary development. Mammary tissue is postulated to undergo epigenetic extensive modelling or re-modelling during different stages in life such as fetal development, puberty or pregnancy. Such epigenetic modification can persist into adulthood if taken place in mammary stem cells, uncommitted mammary myoepithelial or luminal progenitor cells and inherited by subsequent daughter cells [119]. Whether such effects are reversible by later interventions remains to be discovered.

**Table 2-2 Results from a selection of studies investigating different anthropometric measures and risk of adult breast cancer.**

| Age (years) | Anthropometric measure *(increase)* | Effect on breast cancer risk | Remarks | Ref |
|---|---|---|---|---|
| Infant | Birth weight (kg) | ↑ | Meta-analysis of 18 epidemiological studies | [109] |
| Infant | Birth weight (kg) | ↑ | A cohort of 117,415 Danish women | [108] |
| Infant | Birth length (cm) and head circumference (cm) | ↑ | A cohort of 5,358 Swedish women | [110] |
| Infant | Fetal growth rate, as measured by birth size adjusted for gestational age (units/week) | ↑ | A cohort of 5,358 Swedish women | [110] |
| <8 | Change in body mass index (kg/m$^2$) | ↓ | A cohort of 117,415 Danish women | [108] |
| 8-14 | Change in body mass index (kg/m$^2$) | ↓ | A cohort of 117,415 Danish women | [108] |
| 10 | Body mass index (kg/m$^2$) | ↓ | 65,140 women who participated in the Nurses' Health Study | [113] |
| 14 | Body mass index (kg/m$^2$) | ↓ | A cohort of 117,415 Danish women | [108] |
| Young ages | Body fatness (9-level pictogram [level 1: most lean; level 9: most overweight]) | ↓ | A prospective analysis among 188,860 women (7,582 breast cancer cases) | [114] |
| 7-15 | Body mass index (kg/m$^2$) | ↓ | 3,447 Finnish women | [115] |
| Young adult | Body mass index (kg/m$^2$) | ↓ | 10,106 postmenopausal Japanese women | [116] |
| Post-menopause | Body mass index (kg/m$^2$) | ↑ | 10,106 postmenopausal Japanese women | [116] |
| Mean recruitment age 48 years | Body mass index (kg/m$^2$) | ↑ | 424,519 participants from the Asia-Pacific Cohort Studies Collaboration | [120] |

"…*frailty, thy name is woman!*" (Hamlet: I, ii) The complex relationship between body mass and breast cancer risk reverts to a positive association again after a woman ceases to produce hormones naturally in her ovaries (i.e. undergo menopause). There is substantial evidence to support the link between obesity or body mass index or weight gain and breast malignancies in postmenopausal women [116, 120-123]. After menopause, adipose tissue becomes the principle contributor to the circulating pool of estrogen in the body [124]. Estrogen may be implicated in breast cancer risk because it encourages growth of cells in the breast [125].

The effect of adult anthropometric measures on breast cancer risk varies from woman to woman. For example, among women on hormone replacement therapy, thinner women are more likely to get breast cancer than heavier women [126]. On the contrary, among never-users of hormone replacement therapy women with higher BMI was more likely than women with lower BMI to develop breast cancer.

### 2.7.3 Branching into tumour characteristics

But breast cancer is a heterogeneous phenotype – is looking at the overall risk of breast cancer when examining the effects of anthropometric measures enough?

One study by Bardia and colleagues [127] looked into the risk of developing postmenopausal breast cancer stratified by estrogen receptor (ER) and progesterone receptor (PR) subtypes and reported that an increase in weight at age 12 years was associated with a decrease in adult breast cancer risk, with the most pronounced effects exhibited by ER-positive/PR-negative tumours. No significant heterogeneity, however, was observed between the tumour subtypes studied. Adult body mass index, on the other hand, was found to only elevate breast cancer risk for the estrogen receptor positive subtype [128].

# AIMS

The underlying aim of this thesis is to identify common genetic variants that are associated with risk of breast cancer, using both hypothesis-free (Studies I and II) and hypothesis-based (Study III) approaches. To achieve this end, we ventured beyond traditional genetic scans and explored the use of alternative phenotypes (i.e. intermediate phenotype or disease subtype) to see whether the variance explained can be increased. In the last study (Study IV), we look beyond genetics for hints as to why destiny does not always lie in our genes.

**"My lord, I aim a mile beyond the moon"** (Titus Andronicus: IV, iii)

The overarching significance that weaves through all four studies of this research is that, one day, we may:

- ⇒ Classify women according to high or low risk of breast cancer on the basis of genetic disposition and other breast cancer risk factors, so that
- ⇒ Appropriate interventions and disease management decisions may be made, to ultimately
- ⇒ Reduce incidence and mortality of breast cancer.

# 3 MATERIALS AND METHODS

In an attempt to identify common disease susceptibility alleles for breast cancer, we started off with a hypothesis-free approach, and performed a combined analysis of three GWAS, involving 2,702 women of European ancestry with invasive breast cancer and 5,726 controls. Tests for association were performed for 285,984 SNPs.

As GWAS has been said to underperform for studying complex diseases such as breast cancer, we investigated to see if the variance explained by common variants could be increased by studying specific disease subtypes. We performed an independent GWAS using a subset of ER-negative breast cancer cases (N = 617) and all of the controls from the initial genome-wide study.

For both GWAS, we went beyond standard single marker analyses of scan data to look at the importance of groups of SNPs in biologically meaningful pathways using permutation-based tests.

Because mammographic density may be influenced by estrogen, we examined a total of 239 SNPs in 34 estrogen metabolic genes, both on a single marker and global level, in 1,731 Swedish women for associations with mammographic density, which is a strong risk predictor of breast cancer risk.

In addition, even though breast cancers of different ER subtypes are well known to express distinct tumour behaviour and gene expression, it is not known whether they differ in germline genetic risk profiles. The extent of shared polygenic variation between ER-negative and ER-positive breast cancers was assessed by relating risk scores, derived using ER-positive breast cancer samples, to disease state in independent, ER-negative breast cancer cases.

The differential etiology of breast cancers of different ER subtypes was also studied in relation to anthropometric risk factors, such as childhood body size.

## 3.1 SUBJECTS

This thesis made use of subject data from several sources (Table 3-1): breast cases and controls from the Cancer Hormone Replacement Epidemiology in Sweden (CAHRES) study, additional Swedish controls from the Epidemiological Investigation of Rheumatoid Arthritis (EIRA), unselected breast cancer patients and additional familial cases ascertained at the Helsinki University (HUBC), population controls from the Finnish Genome Center (FGC), and cases and controls from the Cancer Genetic Markers of Susceptibility (CGEMS) initiative.

Validation for the genome-wide association scans were performed using the Rotterdam Breast Cancer Study (RBCS) and Studies in Epidemiology and Risks of Cancer Heredity (SEARCH) study, while results of the candidate gene study were validated using subjects from the Mayo Clinic Breast Cancer Study (MBCS) and the Nurses' Health Study (NHS).

**Table 3-1 Summary of data sources used in each study. SNP: single nucleotide polymorphism; ER: estrogen receptor**

| Study | Variable of interest | Outcome of interest | Discovery | Validation |
|---|---|---|---|---|
| *I* | SNPs | Breast cancer | CAHRES<br>EIRA<br>HUBC<br>FGC<br>CGEMS | SEARCH<br>RBCS |
| *II* | SNPs | ER-negative breast cancer | CAHRES<br>EIRA<br>HUBC<br>FGC | SEARCH<br>RBCS |
| *III* | SNPs | Mammographic density | CAHRES | NHS<br>MBCS |
| *IV* | Childhood body size | Breast cancer | CAHRES | -- |

### 3.1.1 Cancer Hormone Replacement Epidemiology in Sweden (CAHRES)

The population-based study, CAHRES, which includes women aged 50-74 years, born in Sweden and resident there between October 1, 1993 and March 31, 1995, is used in all four studies.

An attempt was made to contact all incident cases of invasive primary cancer in this population. Cases were identified through the six Swedish regional cancer registries and were asked to give their written consent to be approached with a mailed questionnaire through their physicians. A total of 3,979 eligible cases were detected of whom 3,345 (84%) participated in the study. Non-participation was due to physcians' refusal (because of psychiatric disorder, anxiety or poor physical health), in 4% and patients' refusal (either to be approached at all or to return to questionnaire or failure in contacting the patient, in 12%. The mean interval from diagnosis to data collection was 4.3 months (standard deviation 1.5 months).

Control women, frequency matched to the expected age distribution of the cases, were randomly selected from a continuously updated Swedish register which provides national registration numbers, name, address and place of birth of each person residing in Sweden. Of 4,188 selected controls, 3,454 (82%) agreed to participate in the study.

Among controls who agreed to participate, 474 (14%) failed to return the mailed questionnaire but subsequently agreed to a telephone interview. No cases were interviewed this way, since 98% of those we had given their consent to receive a questionnaire also returned it. The telephone interview included the most important items in the mailed questionnaire, except family history of breast cancer, weight at age 18, somatotype, menstrual characteristics at age 30, menopausal symptoms and lactation. Controls participating only through the telephone interview did not differ essentially from other controls with regard to the most important risk factors. Approximately 50% of the cases and controls were also contacted by telephone to obtain essential missing information in their mailed responses.

A total of 112 cases and 88 controls, with a previous diagnosis of cancer (other than non-melanoma skin cancer or cancer in situ of the cervix), were excluded. We also excluded pre-menopausal women (198 cases and 152 controls) as well as women with unknown menopausal status (217 cases and 100 controls). The final study population consisted of 2,818 cases and 3,111 controls.

For genetic studies involving DNA specimens in Study III, we sampled eligible women from the parent study described above. We randomly selected 1,500 breast cancer cases among the eligible cases and 1,500 controls that were age-frequency matched to the cases in 5-year intervals. The reason for not including all patients was purely monetary. In addition, all remaining cases (N=301) and controls (N=567) that had taken either medium potency estrogens alone, or estrogen plus progestin preparations for four years or more, were selected. From a total of 1,801 cases and 2,067 controls selected, biological samples from 1,534 cases and 1,504 controls passed quality control for genotyping. This yields approximate population-based participation rates of 84% × 85% = 71% and 82% × 73% = 60% among cases and controls respectively. Of these women, mammograms were available for 891 breast cancer cases and 840 controls.

From the samples selected for genetic studies described above, a subset with sufficient DNA, and information on TNM, lymph nodes, size, grade and outcome, 804 were selected for further genotyping on genome-wide chips (Table 3-2).

**Table 3-2 Completeness of CAHRES data with respect to tumour characteristics**

| Variable | # of samples with information on outcome | # of samples with information on variable on the left |
|---|---:|---:|
| DNA concentration ≥ 35ng/µL | 1,208 | 1,276 |
| and TNM | 1,175 | 1,175 |
| and lymph nodes | 1,174 | 1,178 |
| and tumour size | 1,174 | 1,204 |
| and grade | 804 | 825 |

Out of 804 cases selected for GWAS, one sample could not be matched to phenotype data. Through pairwise clustering in whole genome association analysis software Plink [129], we identified two different pairs of monozygotic twins, one pair on each platform used for genotyping. All four individuals were removed from further analyses as they were most likely the product of a technical mishap. In addition, two pairs of full siblings were found, of which both pairs appeared on both chips. Of these two sibling pairs, the one with the higher call rate was kept for further analyses. A total of 797 cases were included in the GWAS of overall breast cancer risk in Study I. Of these cases, a subset of 153 ER-negative breast cancer cases was selected for GWAS on this particular cancer subtype in Study II.

### 3.1.2 Epidemiological Investigation of Rheumatoid Arthritis (EIRA)

A population-based case–control study on incident cases of rheumatoid arthritis, called EIRA (Epidemiological Investigation of Rheumatoid Arthritis), has been in progress in Sweden since 1996 [130]. The study base comprised the population, aged 18–70 years, living in parts of Sweden during May 1996 to December 2005 [131]. Controls from this study population were used to supplement the Swedish study used in both the overall and ER-negative breast cancer breast cancer GWAS. For each rheumatoid arthritis patient, a control subject was randomly selected from the study base; control subjects were matched for age, sex, and residential area. Most subjects were born in Sweden, and 97% reported having white ancestry.

Exclusions: Nine controls were found to be population outliers by principal component analysis and removed from further analyses.

### 3.1.3 Helsinki University Central Hospital (HUBC)

The Finnish breast cancer study population consists of two series of unselected breast cancer patients and additional familial cases ascertained at the Helsinki University Central Hospital. The first series of patients were collected in 1997-1998 and 2000 and covers 79% of all consecutive, newly diagnosed cases during the collection period [28, 29]. The second series, containing newly diagnosed patients, was collected in 2001 – 2004 and covers 87% of all such patients treated at the hospital during the collection period [30]. The collection of additional familial cases has been described previously [31]. We genotyped a total of 782 breast cancer cases from this study. Of these women, 212 were premenopausal, 359 were postmenopausal, and 211 were missing menopausal status. Population control data was obtained from FGC on 3170 healthy population controls described in [15-18]. A total of 464 ER-negative breast cancer cases, inclusive of an additional 26 sporadic breast cancer patients and 15 *BRCA1* and *5 BRCA2* mutation carriers with ER-negative breast cancer, were used in Study II.

Exclusions: A total of 18 individuals in the Finnish dataset were removed because they were full siblings or monozygotic twins of an individual in the study. In each case, the individual with the highest call rate was kept. In addition, three individuals were removed from the Finnish study population because they were extreme outliers on one or more significant principal component axes. One individual from the Finnish dataset was excluded due to missing affection status.

### 3.1.4 Studies in Epidemiology and Risks of Cancer Heredity (SEARCH)

SEARCH is a population-based case-control study comprising 7,093 cases identified through the East Anglian Cancer Registry: prevalent cases diagnosed age <55 from 1991-1996 and alive when the study started in 1996, and incident cases diagnosed <70 diagnosed after 1996. Controls (N=8,096) were selected from the EPIC-Norfolk cohort study, a population-based cohort study of diet and health based in the same geographical region as SEARCH, together with additional SEARCH controls recruited through general practices in East Anglian region.

### 3.1.5 Rotterdam Breast Cancer Study (RBCS)

RBCS is a hospital-based case-control study comprising 799 cases characterized as familial breast cancer patients selected from the Rotterdam Family Cancer Clinic at the Erasmus Medical Center, of which 141 are ER-negative. Controls

(N=801) were spouses or mutation-negative siblings of heterozygous Cystic Fibrosis mutation carriers selected from the Department of Clinical Genetics at the Erasmus Medical Center. Both cases and controls were recruited between 1994 and 2006.

### 3.1.6 Cancer Genetic Markers of Susceptibility (CGEMS)/ Nurses' Health Study (NHS)

Genotype data was also obtained for a total of 1,145 postmenopausal women of European ancestry with invasive breast cancer from the CGEMS initiative, along with 1,142 matched controls nested within the prospective Nurses' Health Study cohort [16]. The CGEMS project is a National Cancer Institute initiative to conduct genome-wide association studies to identify genes involved in breast cancer and prostate cancer. The initial CGEMS breast cancer scan was designed and funded to study the main effect of SNP variants on breast cancer risk in postmenopausal women, and has been completed. The Nurses' Health Study was initiated in 1976, when 121,700 US registered nurses aged 30 to 55 returned an initial questionnaire [132]. During 1989 and 1990, blood samples were collected from 32,826 women [133]. A subset of 1,590 women - of which 806 were breast cancer cases and 784 were healthy controls - with mammographic density data available were used for the validation of significant SNPs in Study III.

### 3.1.7 Mayo Clinic Breast Cancer Study (MBCS)

The second validation population for Study III consisted of a set of controls from an ongoing breast cancer case-control study at the Mayo Clinic. Briefly, the Mayo Clinic Breast Cancer Study is an Institutional Review Board-approved, clinic-based, case-control study initiated in February 2001 at Mayo Clinic, Rochester, MN, USA. The study design has been presented previously [15, 134]. Clinic attendance formed the sampling frame for Mayo Clinic cases and controls. Consecutive cases were women aged 18 years or over with histologically confirmed primary invasive breast carcinoma and recruited within 6 months of the date of diagnosis. Cases lived in the six-state region that defines Mayo Clinic's primary service population (Minnesota, Iowa, Wisconsin, Illinois, North Dakota, and South Dakota). Controls without prior history of cancer (other than nonmelanoma skin cancer) were frequency matched on age (5-year age category), race and six-state region of residence to cases. Controls were recruited from the outpatient practice of the Divisions of General Internal Medicine and Primary Care Internal Medicine at Mayo Clinic, where they were seen for routine medical examinations.

The analysis in Study III was performed on genotyped Caucasian controls (99% of study participants) enrolled through September 2007, who had mammograms available, representing 995 total controls (76% of total possible controls), of which 783 were postmenopausal.

For all populations, blood samples were obtained from individuals according to protocols and informed-consent procedures approved by institutional review boards.

## 3.2 DATA COLLECTION

### 3.2.1 Key variables

*Genotypes*. Genotyping for all samples in Studies I-III was performed according to the manufacturers' instructions. Table 3-3 below summarizes the different platforms used for genotyping.

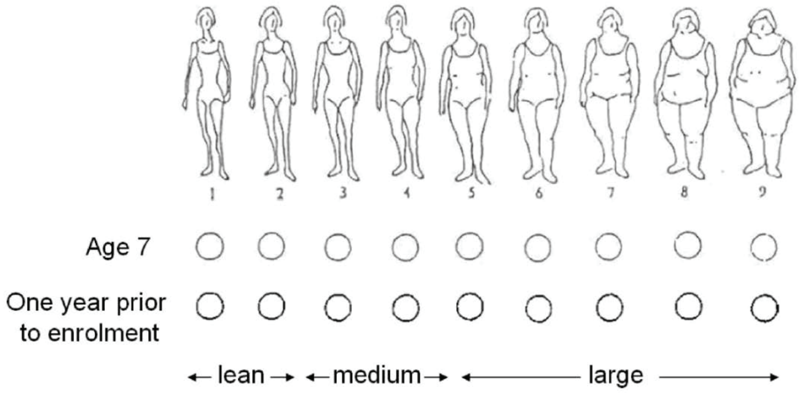
**Table 3-3 Summary of the different platforms used for genotyping**

| Paper | Population | Platform |
|---|---|---|
| I/III | CGEMS/NHS | Illumina Infinium 2 HumanHap550 |
| I/II | CAHRES (cases) | Illumina Infinium 2 HumanHap300, HumanHap240S |
| I/II | CAHRES (controls) | Illumina Infinium 2 HumanHap550 |
| I/II | EIRA | Illumina Infinium 2 HumanHap300 |
| I/II | HUBC (cases) | Illumina Infinium 2 HumanHap550 |
| I/II | Finnish Genome Centre | Illumina Infinium 2 HumanHap370Duo |
| I/II | SEARCH | Taqman |
| I/II | RBCS | Taqman |
| III | CAHRES | Sequenom |
| III | MBCS | Taqman |

*Somatotypes*. Anthropometric measurements at age seven years and one year prior to enrolment in Study IV were collected by means of a nine-level somatotype (Figure 3-1) featured in the study questionnaire. The somatotypes were subsequently grouped as lean (S1 to S2), medium (S3 to S4) and large (S5 to S9) prior to analysis.

Breast cancer patients were identified at diagnosis through the six Swedish regional cancer registries, to which the reporting of all malignant tumors is mandatory. The “personnummer” or personal number is a unique national identity number unique to all Swedish residents. The date of birth and the sex of an individual may be easily derived from the personnummer, and it is the key in most government databases. It is possible for researchers, provided that the appropriate permissions are granted, to approach the authority in charge of the Total Population Register (currently known as the Tax Authority) and ask for the national registration numbers and addresses of people that fulfill certain criteria specified by the researcher.

**Figure 3-1 Nine-level somatotype pictogram**



Information regarding the retrieval of tumour characteristics from the medical records of all participants from surgical and oncological units throughout Sweden has been presented in detail elsewhere [135, 136]. The tumour characteristics in the present study included tumour size (categorical, groups in cm), grade (categorical, classified according to the Nottingham histological grade or Bloom-Richardson scale), as well as ER and PR status (binary, absent/present).

Breast tumours were routinely measured for ER and PR content in Sweden at the time of the study, but the assessment was often not performed on small tumours (≤1 cm in size) due to lack of material. Receptor analyses were usually performed by one laboratory within each region. All seven laboratories in Sweden analyzing ER and PR content used an enzyme immuno-assay (Abbott Laboratories) on cytosol samples. The method used for assessing ER content was ERα specific [137]. Three laboratories reported amount of receptor per μg DNA, three laboratories reported amount of receptor per mg protein, and one laboratory reported both. Quantitative receptor content was available for 67% of the tumours for both ER and PR. In 4% and 3% of the tumours, for ER and PR, respectively, instead of quantitative data, information on tumour status was classified as being strongly positive, positive, weakly positive, or negative instead. Receptor positive tumours as defined as ≥0.05 fmol receptor/μg DNA or ≥10 fmol receptor/mg protein. For the laboratory reporting both analyses, the proportion of receptor positive tumours was most similar to the proportion among the other laboratories when measured as amount of receptor per μg DNA. Hence, we used these values. Tumours with qualitative information were defined as receptor negative if they were classified as negative, otherwise they were classified as positive.

The process of collecting mammographic density data in this study has been described previously [138]. Film mammograms of the medio-lateral oblique view were digitized using an Array 2905HD Laser Film Digitizer (Array Corporation, Tokyo, Japan), which covers a range of 0 to 4.7 optical density. For controls, the breast side was randomized. For cases, the side contralateral to the tumor was used. The density resolution was set at 12-bit spatial resolution. The Cumulus software used for the computer-assisted measurement was developed at the University of Toronto [139]. For each image, a trained observer (Louise Eriksson) set the appropriate gray-scale threshold levels defining the edge of the breast and distinguishing dense from nondense tissue. The software calculated the total number of pixels within the entire region of interest and within the region identified as dense. These values were used to calculate the percentage of the breast area that is dense. A random 10% of the images were included as replicates to assess the intra-observer reliability, which was high with a Spearman rank correlation coefficient of 0.95.

## 3.3 Statistical analyses

The analytical strategy for the two genome-wide studies (Studies I and II) is summarized in Figure 3-2. To assess the enrichment of significant associations in a pathway context, two different pathway analysis tools were used: SNP Ratio Test (SRT) [140] and admixture maximum likelihood (AML) [141] test. The "scoring" approach by Purcell et al. [142] was used to assess the extent of common polygenic variation between ER-positive and ER-negative breast cancers in Study II. The Plink [129] option "--adjust" was used to generate a file of adjusted significance values that correct for all tests performed and other metrics in Studies I and II. To reduce the impact of population stratification on the genome-wide studies, principal component analysis (PCA) [143] was performed.

Regression models were used in all studies to predict the outcome from one or more independent variables. For a dichotomous outcome such as breast cancer case/control status, logistic regression models were fitted. For a continuous outcome such as percent mammographic density, linear regression models were fitted. For ordinal outcomes with more than two values such as body size, grade or tumour size categories, proportional odds logistic regression (POLR) models were fitted. A summary of methods and corresponding software for analysis pertaining to all four studies is presented in Table 3-4.

Information was gathered from several public databases (Table 3-5). Genotype data on the Cancer Genetic Markers of Susceptibility (CGEMS) [16] data set was obtained with permission from the Database of Genotypes and Phenotypes (dbGaP). Candidate SNPs were annotated using the SNP and CNV Annotation Database (SCAN) [144]. A web-based tool called POLYSEARCH [145] was used to mine information in published literature on relationships between novel candidate genes and disease. Pathway definitions were obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG) [146], and a list of SNPs regulating gene expression was downloaded from mRNA by SNP Browser [57]. Pairwise LD between SNPs were obtained from SNP Annotation and Proxy Search (SNAP) [147].

Data visualization was enabled using the following freeware: HaploView [148], LocusZoom [149], Edraw Mindmap [150], and R [151]. Details on what the different programs were used for are listed in Table 3-6. A list of other tools used in this thesis is available in Table 3-7.

**Figure 3-2 Schematic diagram of analytical strategies for agnostic single marker association analysis and pathway analysis.**

The study designs for Studies I and II differ in i) the study populations used and ii) specifics of pathway analysis. Study I consists of genotype data from three independent breast cancer case-control populations – Swedish, Finnish and American (CGEMS); Study II consists only of ER-negative breast cancer cancers and controls from the Swedish and Finnish populations. Study II has an additional pathway analysis performed using pathway definitions from SNPs associated with gene expression.

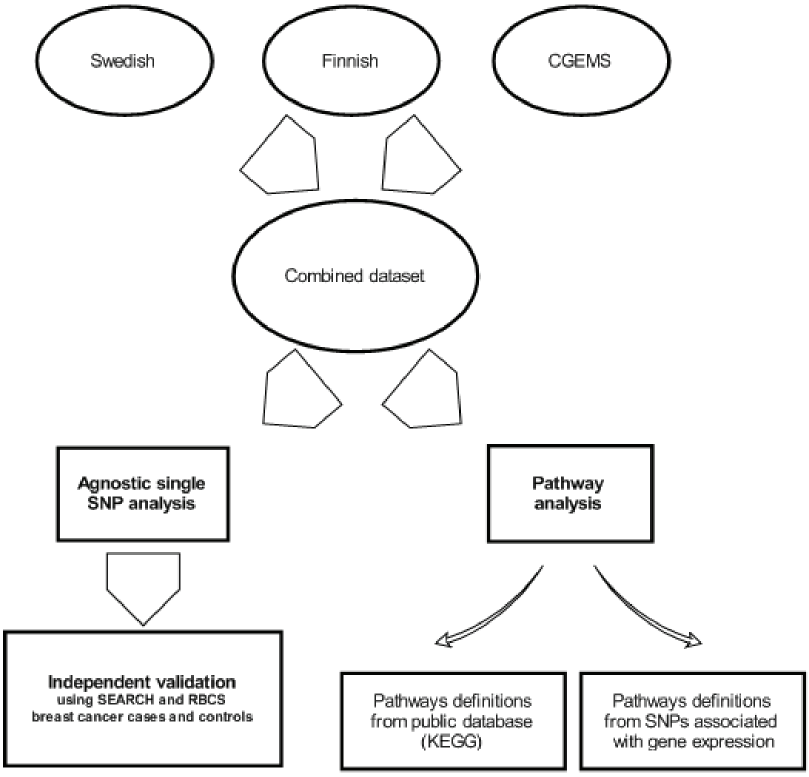

Swedish
Finnish
CGEMS
Combined dataset
Agnostic single SNP analysis
Pathway analysis
Independent validation using SEARCH and RBCS breast cancer cases and controls
Pathways definitions from public database (KEGG)
Pathways definitions from SNPs associated with gene expression

**Table 3-4 Summary of methods and corresponding software for analysis**
BBD: benign breast disease; BMI: body mass index; Meno: menopausal status; HRT: hormone replacement therapy; PC: principal component; GC: genomic control; SRT: SNP Ratio Test; AML: admixture maximum likelihood

| *Method* | *Independent variable* | *Dependent variable* | *Adjusted variables* | | | | | | | | *Software* | *Paper(s)* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **Age** | **Age of menarch** | **BBD** | **BMI** | **Meno** | **HRT** | **PC** | **GC** | | |
| Logistic regression | Genotype (0, 1 and 2) | Disease status (case/control) | | | | | | | ● | | Plink [129] | I, II |
| | Somatotype at age 7 (categorical) | Disease status (case/control) | ● | ● | ● | ● | | | | | R [151] | IV |
| Linear regression | Genotype (0, 1 and 2) | Mammographic percent density (%) | ● | | | ● | ● | ● | | | R [151] | III |
| Proportional odds logistic regression (POLR) | Somatotype at age 7 (categorical) | Ordinal outcomes (e.g. grade, tumour size) | ● | | | | | | | | R [151] | IV |
| SNP set enrichment analysis (Pathway analysis) | SNP-set genotypes | Disease status (case/control) | | | | | | | ● | | SRT [140] | I, II |
| | | | | | | | | | | ● | AML [141] | II |
| | | Lowest vs highest mammographic percent density quantiles | | | | | | | | | | III |
| Assessment of common polygenic variation | Scores | Disease status (case/control) | | | | | | | ● | | Plink [129] | II |
| Correction for multiple testing | -- | -- | | | | | | | | | QVALUE [152, 153], Plink | I |
| Principal component analysis (PCA) | -- | -- | | | | | | | | | EIGENSTRAT [143] | I, II |

**Table 3-5 List of online databases**

| *Database* | *Link* | *Remark* | *Paper(s)* |
|---|---|---|---|
| Cancer Genetic Markers of Susceptibility (CGEMS) [16] | cgems.cancer.gov | CGEMS genotypes | I |
| SNP and CNV Annotation Database (SCAN) [144] | scan.bsd.uchicago.edu/newinterface/about.html | A large-scale database of genetics and genomics data for annotating candidate SNPs | I |
| POLYSEARCH [145] | wishart.biology.ualberta.ca/polysearch/ | Database for biomedical text mining | I |
| Kyoto Encyclopedia of Genes and Genomes (KEGG) [146] | www.genome.jp/kegg | Pathway definitions | I, II |
| mRNA by SNP Browser [57] | www.sph.umich.edu/csg/liang/asthma | Database of genetic variants regulating gene expression | II |
| SNP Annotation and Proxy Search (SNAP) [147] | http://www.broadinstitute.org/mpg/snap/ | Information on pairwise LD | Kappa |

**Table 3-6 List of graphic making tools**

| *Software* | *Purpose* | *Paper(s)* |
|---|---|---|
| HaploView [148] | Manhattan plot | I, II |
| LocusZoom [149] | Regional visualization of genome-wide association scan results | II |
| Edraw Mindmap [150] | Flowchart | II |
| R [151] | Quantile-quantile plot | I-III |

**Table 3-7 List of other tools**

| *Software* | *Purpose* | *Paper(s)* |
|---|---|---|
| Quanto [154] | Power calculation | II |
| Qlikview (v8.5) [47] | Map SNPs, which are significantly associated with gene expression on a genome-wide level (LOD>6), to genes | II |

# 4 RESULTS

In this section, I present results from individual articles supplemented by data which is neither shown in the main article nor the supplementary materials of the four published manuscript.

## 4.1 STUDY I

In Study I [155], we confirmed associations with loci reported by previous GWAS on 1p11.2, 2q35, 3p, 5p12, 8q24, 10q23.13, 14q24.1 and 16q (Table 4-1), and we presented evidence that suggested novel SNPs for overall breast cancer risk, but the associations from our combined GWAS were not validated in the independent SEARCH and RBCS studies (Table 3 in Article 1). We also assessed evidence for association with SNPs in genes in specific pathways using permutation-based tests, but found only suggestive evidence. Androgen and estrogen metabolism, a pathway previously found to be associated with the development of postmenopausal breast cancer, was marginally significant with overall breast cancer risk (P = 0.084).

**Table 4-1 Pair-wise linkage disequilibrium is pre-calculated based on phased genotype data from the International HapMap Project (Release 22).**
LD data is calculated using a pairwise LD tool called SNP Annotation and Proxy Search (SNAP), created by the Broad Institute. The panel consisting of Utah residents with ancestry from northern and western Europe was used. r2 is a measure of LD which ranges between 0 (when they are in perfect equilibrium) and 1 (when the two markers provide identical information). Study SNP denotes the SNP in Study I that was used for comparison with a corresponding established breast cancer SNP. P denotes the combined P value per allele of three studies.

| PMID | SNP | CHR | BP | GENE | Study SNP | R2 | P |
|---|---|---|---|---|---|---|---|
| 19330030 | rs11249433 | 1 | 120982136 | INTERGENIC | rs11249433 | 1.000 | 1.13E-05 |
| 17529974 | rs13387042 | 2 | 217614077 | INTERGENIC | rs13387042 | 1.000 | 9.23E-06 |
| 19330027 | rs4973768 | 3 | 27391017 | *SLC4A7* | rs4973768 | 1.000 | 1.41E-04 |
| 18438407 | rs10941679 | 5 | 44742255 | INTERGENIC | rs7716600 | 0.784 | 7.06E-07 |
| 18438407 | rs4415084 | 5 | 44698272 | INTERGENIC | rs4415084 | 1.000 | 1.74E-04 |
| 17529967 | rs889312 | 5 | 56067641 | INTERGENIC | | | |
| 19219042 | rs2046210 | 6 | 151990059 | INTERGENIC | | | |
| 20453838 | rs3757318 | 6 | 151955806 | *C6orf97* | | | |
| 17529967 | rs13281615 | 8 | 128424800 | INTERGENIC | rs672888 | 0.967 | 5.29E-05 |
| 20453838 | rs1562430 | 8 | 128457034 | INTERGENIC | rs672888 | 0.426 | 5.29E-05 |
| 20453838 | rs1011970 | 9 | 22052134 | INTERGENIC | | | |
| 20453838 | rs10995190 | 10 | 63948688 | *ZNF365* | | | |
| 17529973 | rs1219648 | 10 | 123336180 | *FGFR2* | | | |
| 20453838 | rs2380205 | 10 | 5926740 | INTERGENIC | | | |
| 17529967 | rs2981582 | 10 | 123342307 | *FGFR2* | rs1219648 | 1.000 | 1.93E-13 |
| 19536173 | rs2981582 | 10 | 123342307 | *FGFR2* | rs1219648 | 1.000 | 1.93E-13 |
| 19536173 | rs3135718 | 10 | 123343859 | *FGFR2* | | | |
| 20453838 | rs704010 | 10 | 80511154 | INTERGENIC | | | |
| 19536173 | rs7895676 | 10 | 123323987 | *FGFR2* | Query SNP not in Release 22 | | |
| 17529967 | rs3817198 | 11 | 1865582 | *LSP1* | | | |
| 20453838 | rs614367 | 11 | 69037945 | INTERGENIC | | | |
| 20453838 | rs909116 | 11 | 1898522 | *TNNT3* | | | |
| 19330030 | rs999737 | 14 | 68104435 | *RAD51L1* | rs999737 | 1.000 | 8.30E-03 |
| 17529967 | rs12443621 | 16 | 51105538 | *TOX3* | rs3803662 | 0.332 | 4.06E-07 |
| 17529967 | rs3803662 | 16 | 51143842 | *LOC643714* | rs3803662 | 1.000 | 4.06E-07 |
| 17529974 | rs3803662 | 16 | 51143842 | *LOC643714* | rs3803662 | 1.000 | 4.06E-07 |
| 17529967 | rs8051542 | 16 | 51091668 | *TOX3* | No LD data is available for rs8051542 in Release 22 (CEU) | | |

In an independent pathway analysis of breast cancer GWAS, Menashe et al. [156] examined a total of 421 pathways retrieved from three databases, of which 155 belong to KEGG [146]. The smallest p-value associated with a KEGG pathway obtained was for basal cell carcinoma (P = 0.00463), in which 53 genes were considered. Table 4-2 shows juxtaposed results of the most associated pathways (P < 0.1) in Study I and the corresponding results from Menashe et al.. The smallest p-value obtained for the latter data set was 0.084 corresponding to the alpha-linolenic acid metabolism pathway.

**Table 4-2 A comparison of the overlapping results of 182 KEGG pathways examined in Study I and 155 KEGG pathways examined by Menashe et al. (141).**

* The other glycan degradation pathway has got two sub-pathways, of which Menashe et al. analyzed independently: N-glycan and glycosphingolipid biosynthesis – ganglioseries respectively.

| | Li et al. | | | Menashe et al. | |
|---|---|---|---|---|---|
| *Name* | *# SNPs P < 0.05* | *# SNPs in pathway* | *P* | *# genes* | *P* |
| Other glycan degradation | 11 | 62 | 0.004 | 44 | 0.533* |
| | | | | 16 | 0.531* |
| Pentose and glucuronate interconversions | 9 | 62 | 0.035 | 25 | 0.400 |
| **alpha-Linolenic acid metabolism** | **11** | **71** | **0.010** | **15** | **0.084** |
| Drug metabolism - other enzymes | 25 | 227 | 0.036 | 49 | 0.913 |
| Inositol phosphate metabolism | 44 | 584 | 0.054 | 47 | 0.174 |
| Androgen and estrogen metabolism | 19 | 199 | 0.084 | 52 | 0.991 |
| Hematopoietic cell lineage | 36 | 404 | 0.010 | 79 | 0.213 |
| Primary immunodeficiency | 10 | 111 | 0.082 | 34 | 0.367 |
| Regulation of actin cytoskeleton | 129 | 1870 | 0.012 | -- | -- |
| Circadian rhythm | 13 | 89 | 0.014 | 13 | 0.605 |
| Thyroid cancer | 20 | 217 | 0.057 | 29 | 0.587 |

## 4.2 STUDY II

The general analytical strategy of Study II was similar to that of Study I – genome-wide association analysis on a single marker and pathway level – but on a different phenotype. Association with ER-negative breast cancer was not validated for any of the five most strongly associated SNPs followed up in independent studies (1,011 ER-negative breast cancer cases, 7,604 controls) (Table S2 in Additional File 3 of Article 2). However, in this study, we presented additional pathway analysis results based on a selection of pathways defined by an exclusive set of SNPs reported to be associated with gene expression (Table 3 in Article 2). An excess of small p-values for SNPs with known regulatory functions in cancer-related pathways was also found (global P = 0.052, Figure 9 in Article 2). In addition, we found no evidence to suggest that ER-negative breast cancer shares a polygenic basis with ER-positive breast cancer (Figure 10 in Article 2).

Encouraged by the signal peak observed in the Manhattan plot of association results with ER-negative breast cancer risk on chromosome 9 (Figure 3 in Article 2), and the strong associations with SNPs rs7039994 and rs12000794 within the region (p-values of $3.95 \times 10^{-06}$ and $5.40 \times 10^{-06}$, respectively), our collaborators at the Department of Biosciences and Nutrition carried out preliminary fine-mapping and functional studies on the *INVS* gene in parallel with the validation in the SEARCH/RBCS samples (data not shown in Article 2). In a test data set consisting of 174 ER-negative cases and 325 controls (108 ER-negative cases and

158 controls overlapping with the samples used for GWAS), 23 SNPs, which include missense coding SNPs, promoter SNPs and tagging SNPs within the within the region of the INVS gene, but not found on the SNP array, were genotyped at the Mutation Analysis Facility (MAF) at Karolinska Institutet, Huddinge, Sweden (www.maf.ki.se) using matrix-assisted laser desorption/ ionization time of-flight (MALDI-TOF) mass spectrometry (Sequenom, San Diego, CA, USA). A total of 15 SNPs that were successfully genotyped, and which were not monomorphic in the European population were tested for association with breast cancer risk.

Table 4-3 shows the association results for ER-negative breast cancer risk in the partially overlapping data set - rs7039994 and rs12000794 were not found to be significantly associated at the 5% level (P = 0.1446 and P = 0.2496, respectively). Three SNPs were found to be marginally associated: rs2209081, rs2806684 and rs17812397.

**Table 4-3 Association results of breast cancer risk performed on 15 additional SNPs within the INVS gene region, using 174 ER-negative cases and 325 controls**

There is an overlap of 108 ER-negative cases and 158 controls with the CAHRES GWAS data set. *SNPs rs7039994 and rs12000794 are also found on the GWAS panel.

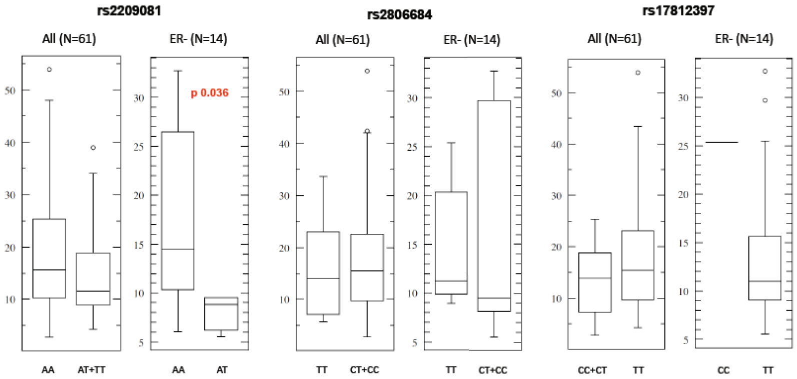
| | | MAF | | |
|---|---|---|---|---|
| **SNP** | **Allele** | **Cases** | **Controls** | **P** |
| rs6479002 | G | 0.445 | 0.399 | 0.1665 |
| rs62577237 | T | 0.018 | 0.016 | 0.7855 |
| rs7024375 | T | 0.839 | 0.835 | 0.8868 |
| rs2209081 | A | 0.793 | 0.733 | **0.0374** |
| rs4273907 | C | 0.921 | 0.899 | 0.2680 |
| rs12003061 | A | 0.185 | 0.155 | 0.2377 |
| rs2806684 | T | 0.618 | 0.552 | **0.0481** |
| rs41312220 | A | 1.000 | 0.994 | 0.1451 |
| rs2787390 | G | 0.891 | 0.888 | 0.9049 |
| rs17812397 | C | 0.166 | 0.119 | **0.0436** |
| rs875522 | A | 0.276 | 0.235 | 0.1636 |
| rs2787371 | C | 0.408 | 0.398 | 0.7705 |
| *rs7039994 | T | 0.252 | 0.198 | 0.1446 |
| rs7029342 | T | 0.225 | 0.195 | 0.2831 |
| rs10123866 | G | 0.734 | 0.698 | 0.2479 |
| rs7048601 | G | 0.500 | 0.490 | 0.7714 |
| *rs12000794 | A | 0.272 | 0.227 | 0.2496 |

.

We next looked at *INVS* expression by genotype in tumour tissue (data not shown in Article 2). Total RNA was extracted from 61 breast cancer tumors, of which 14 were ER-negative, using RNeasy fibrous mini kit (Qiagen, Hilden, Germany), with small modifications from the manufacturer's instructions. The cDNA was synthesized using SuperScript™ III Reverse Transcriptase reagents (Invitrogen, Carlsbad, CA, USA), according to manufacturers' instructions. TaqMan assays specific for *INVS* (Hs00205297_m1) were used for the qRT-PCR (Applied Biosystems, Foster City, CA, USA), and performed in triplicates with the 7500 Fast Real-Time PCR system using standard protocols (Applied Biosystems, Foster City, CA) with a final sample volume of 10µ. Cycle threshold (CT) values were analyzed and obtained using 7500 SDS software (Applied Biosystems). A TaqMan assay for the endogenous housekeeping gene *GAPDH* (Hs99999905_m1) (Applied Biosystems) was used for normalization and relative quantification.

Among the three SNPs found to be significantly associated to ER-negative breast cancer risk via additional genotyping, no significant difference in gene expression was found when both ER-positive and ER-negative tumours were considered together (Figure 4-1). However, when only ER-negative breast cancer tumours were considered, *INVS* expression was found to be significantly higher (P = 0.036) in homozygous risk allele carriers.

**Figure 4-1 Results from gene expression study**



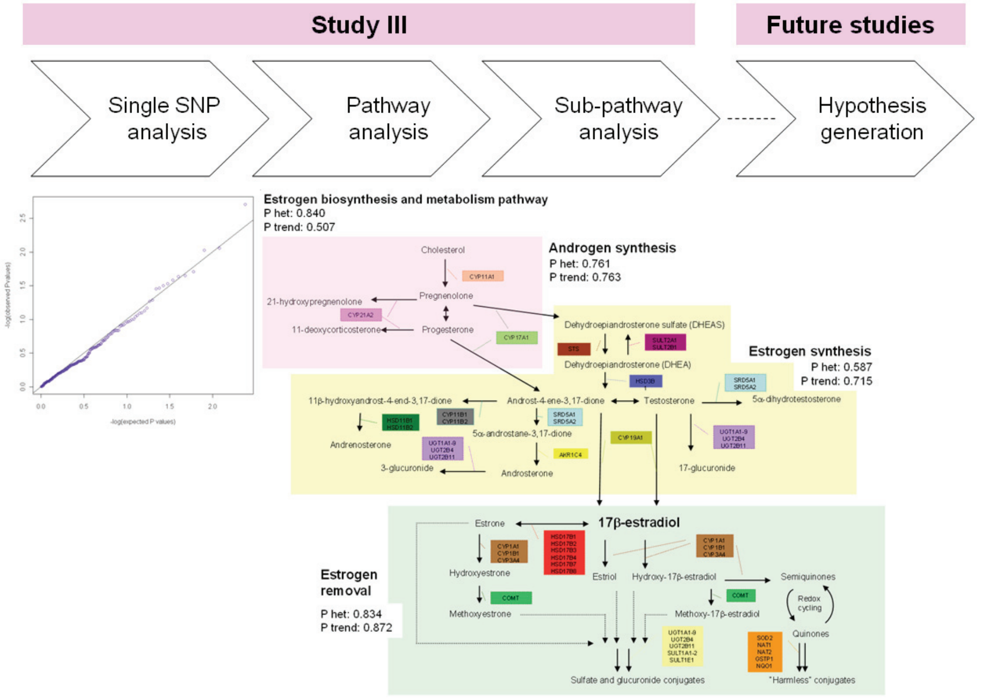
## 4.3 STUDY III

We examined 239 SNPs in the estrogen metabolic pathway for associations with mammographic density on three different levels: single SNP, overall estrogen metabolic pathway, and its sub-pathways. Figure 4-2 shows a summary of the different levels of analysis and corresponding results performed in Study III. Quantile-quantile P value plots from single-SNP trend tests of 239 SNPs in the estrogen metabolism pathway showed no clear deviation from the null distribution, representing no association between SNPs and percentage density. None of the SNPs found to be nominally significant in our dataset were found to be significant in either the CGEMS or MBCS validation sets. Pathway-based multi-SNP association analyses revealed no significant association between percentage density and genetic variations in the overall estrogen metabolic pathway, or any of the related sub-pathways.

Since the estrogen metabolic SNPs examined have previously been associated with breast cancer risk, we also estimated the correlation between regression coefficients of SNP effects on mammographic density and the odds ratios of SNP effects on breast cancer risk, in order to assess whether the SNPs act through mammographic density as an intermediate phenotype for breast cancer. No significant relationship was found between SNP effects on breast cancer risk and percentage density (Spearman's correlation rho = 0.0411, P = 0.5268).

Although percent mammographic density is strongly associated with hormone-related factors, we did not find a conclusive association between any genetic variants in the estrogen metabolic pathway examined and percent mammographic density, although we performed exhaustive analyses [157].

**Figure 4-2 Summary of the different levels of analysis and corresponding results performed in Study III.**
Quantile-quantile P value plots from single-SNP trend tests of 239 SNPs in the estrogen metabolism pathway showed no clear deviation from the null distribution, representing no association between SNPs and percentage density. Pathway-based multi-SNP association analyses revealed no significant association between percentage density and genetic variations in the overall estrogen metabolic pathway, or any of the related sub-pathways.

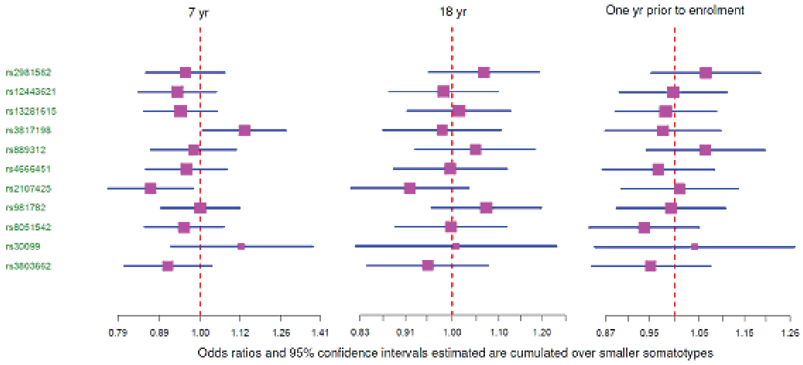


## 4.4 STUDY IV

In Study IV we studied the risks of ER subtype breast cancers with respect to childhood body size [158]. We found that a larger somatotype exerts a larger protective effect in ER-negative than ER-positive breast cancers. The significant protective effect was observed within all subgroups defined by estrogen receptor (ER) and progesterone receptor (PR) status, with a stronger effect for ER-negative (0.40, 95% CI = 0.21-0.75, P trend = 0.002), than for ER-positive (0.80, 95% CI = 0.62-1.05, P trend = 0.062), tumours (P heterogeneity = 0.046). Somatotype at age 7 was not associated with tumour size, histology, grade or the presence or absence of metastatic nodes.

Given the strong association with breast cancer risk, we also examined the role of genetic variation in body size development during childhood, adolescence, and adulthood.

**Figure 4-3 Effects of breast cancer susceptibility SNPs on somatotypes at age 7, age 18, and one year prior to enrolment**



# 5 DISCUSSION

## 5.1 STUDIES I AND II

In our combined GWAS of overall breast cancer risk, we confirmed associations with loci reported by previous GWAS on 1p11.2, 2q35, 3p, 5p12, 8q24, 10q23.13, 14q24.1 and 16q, but did not validate any new associations (Figure 4-1). We also assessed evidence for association with SNPs in genes in specific pathways using permutation-based tests, but found only suggestive evidence. Androgen and estrogen metabolism, a pathway previously found to be associated with the development of postmenopausal breast cancer, was marginally significant for an association with breast cancer. In a candidate gene study performed on a largely independent set of SNPs within the estrogen metabolism pathway involving the CAHRES subjects (1,596 breast cancer cases and 1,730 controls) [159], we found corroborating evidence of association with breast cancer risk (global P = 0.034). Further testing revealed the association to be focused on polymorphisms within the androgen-to-estrogen conversion sub-pathway (global P = 0.008). Further tumour subtype analysis demonstrated that the association of the sub-pathway with breast cancer risk was confined to estrogen receptor positive tumours (global P = 0.0003). These results suggest that further analysis of SNPs in these pathways may identify associations that would be difficult to detect through agnostic single SNP analyses.

We also performed an independent GWAS using a subset of ER-negative breast cancer cases (N = 617) and all of the controls from the initial GWAS. We found an excess of SNPs which were linked to gene expression, and significantly associated to breast cancer, within cancer-pathways defined by KEGG. In addition, we also demonstrated that ER-negative cancers only share a fraction of their polygenic component with ER-positive breast cancers.

### 5.1.1 Population stratification

Population stratification is a potential source of bias in GWAS [160]. Uncorrected population structure in an association study may lead to spurious associations and compromise power to detect real associations [161]. A correction for population stratification may be performed by identifying and assigning each subject to a subpopulation (Finnish, Swedish, CGEMS, for example), or by finding underlying principal components (PCs) [143]. Population stratification was addressed in the logistic regression model by including a variable denoting the country of origin and relevant PCs for each population for the combined analyses in Studies I and II. The pros of principal component analysis (PCA) are that it is computationally fast, provides information about population substructure, and enhances the power of association analysis. On the other hand, one of the cons is that direct application of PCA may not work when related samples or ambiguous relationships are present, thus care was taken to remove all relations and outliers in our GWAS data sets prior to the application of PCA.

### 5.1.2 Imputation

In a GWAS, there are often SNPs which are not genotyped due to reasons such as genotyping errors. Missing SNP data is not uncommon in association studies, sometimes with rates as high as 5-10% [162]. Unfortunately, re-genotyping is usually not possible due to financial constraints. Redundancy of information due to linkage disequilibrium and cost of genotyping are other justifications for not genotyping all available SNPs in the genome.

In genetics, we can deal with the missing data points by substituting missing data values by a method called imputation, which is usually performed by analyzing all existing values to determine the most likely value for the missing data points. Imputation of missing SNPs is made possible with the International HAPMAP Project [163] and other data deposits of genome-wide LD structure in different populations. Internal imputation refers to the imputation of SNPs which are missing in a subset of individuals in genotyped SNPs using LD information in observed data. External imputation refers to the imputation of a large number of untyped SNPs using external LD information from HAPMAP, for example. As long as the error is not too large, imputation may increase statistical power in a GWAS, because it gives more marker information (see sub-section on *Increase resolution* on p18). However, if genotype information for cases and controls are obtained from different chips, as was the case for our GWAS data sets, false-positive associations may arise from the imputation of untyped SNPs in both cases and controls from a haplotype panel due to a difference in imputation quality [164]. In addition, imputation rarely increases power beyond that of actually genotyped SNPs [165], and imputation is often tricky for uncommon variants [166]. The decision was made not to perform imputation on our data sets upon weighing the limited potential gains [167] and practical aspects of an imputation-driven meta-analysis of GWAS [168].

### 5.1.3 Pathway analysis

For a polygenic disease such as breast cancer, it is possible that genetic contributions operate through numerous SNPs in multiple genes in functional pathways. Several studies have taken a pathway analysis approach to elucidate these mechanisms using data from high resolution SNP arrays [156, 169]. However, the nascent interest in pathway analysis on GWAS data means that a standard workflow of strict guidelines or best practices does not yet exist, and new approaches arise every day [170, 171].

**Matching SNPs to genes.** There is not yet a consensus in the way each gene within a pathway is represented by SNPs. For example, Menashe et al. [156] reported a genome-wide pathway analysis study on breast cancer risk where a total of 421 pathways containing 3,962 genes retrieved from three databases were examined. Among the results for the common pathways defined by KEGG, only one pathway was found to rank highly in both studies: alpha-Linolenic acid metabolism (P in Study I = 0.01, P in Menashe et al. [156] = 0.084) (Table 4-2). The lack of corroborating evidence could be due to the fact that while all SNPs within each gene were considered in Study I, only the most strongly associated SNP was used to represent each gene in Menashe et al..

We demonstrated the same ambiguity of using SNPs to represent genes in Study II. Other than considering SNPs located within transcript of genes as a set for pathway analysis, we also looked at only SNPs that had a direct association with expression levels for genes within each pathway. While the pathway analysis results for the latter were mostly borderline and suggestive, we found an excess of SNPs which were linked to gene expression, and significantly associated to breast cancer, within cancer-pathways defined by KEGG.

**Matching genes to pathways.** An additional limitation of the pathway-based approach for GWAS analysis is that the definition of a pathway is not always clear. Proprietary and public databases such as MetaCore [172], KEGG [146], BioCarta [173], or the NCI Pathway Interaction Database [174] are collections of manually drawn maps representing the most recent knowledge on the molecular interaction and reaction networks, and could be subject to human error. As definitions are often not cross-checked among databases, there is an extensive amount of overlap and redundancy, making it difficult to use in a genome-wide setting.

Pathway analysis comes with a *caveat*: that it is heavily based on numerous assumptions for every step and the results may turn out quite different depending on how the analysis is run. But the beauty of a pathway analysis is that the same SNPs need not come up across replication sets – it would suffice so long the same pathway tops the list. For example, there is little overlap between the 239 SNPs in the estrogen metabolic pathway in Study III and the 199 SNPs in the androgen and estrogen metabolism pathway defined by KEGG in Study I. However, the related pathways were both found to be associated with breast cancer risk.

A debate on whether pathway analysis is a viable analytical approach opens up a giant can of worms. Despite the many outstanding issues regarding pathway analysis in GWAS (e.g. defining pathways and genes using SNPs correctly), its application holds great promise for making headway in understanding breast cancer etiology beyond the level of single markers.

## 5.2 STUDY III

In Study III, we evaluated a total of 239 SNPs in 34 genes in the estrogen metabolic pathway in 1,731 Swedish women who participated in a breast cancer case-control study. Of the 1,731 women in this study, 891 were cases and 840 were controls. Despite a large sample size and the most complete coverage of the estrogen metabolic pathway, the results led us to conclude that there is no appreciable evidence that genetic variants in genes involved in the estrogen metabolic pathway are associated with mammographic density in postmenopausal women.

Other linkage and candidate gene association studies have also been largely unsuccessful in identifying loci related to mammographic density [69]. The failure to find more genes is, in retrospect, unsurprising since the total number of candidate genes evaluated is only a small proportion of the total number of genes in the human genome. As it has been demonstrated that population variation in percent mammographic density at a given age has high heritability [22], a genome-wide study may offer clues on which variants are responsible for determining mammographic density. To this end, using a subset of the CAHRES data with both GWAS and mammographic density information, we joined four groups with similar data sets to form the Marker of DEnsity (MODE) consortium. A meta-analysis of the five GWAS, consisting of a total of 4,887 women, was conducted for percent mammographic density adjusted for age and BMI, and we found a novel association between rs10995190 in *ZNF365* and percent mammographic density (manuscript accepted for publication in Nature Genetics). The same SNP was recently identified in a GWAS as a breast cancer susceptibility locus [77], suggesting that one or more variants in *ZNF365* could affect breast cancer risk by influencing the proportion of dense tissue in the breast.

## 5.3 STUDY IV

In an epidemiologic study (Study IV) examining the relationship between childhood body size and breast cancer risk, we found that greater body size at age 7 is associated with a decreased risk of postmenopausal breast cancer, and that the associated protective effect is stronger for the ER-negative breast cancer subtype than for the ER-positive subtype. We speculate that body size at age 7 could have permanent repercussions on the epigenetic level that persists to adult life to influence the risk of postmenopausal breast cancer. When considering a genetic aspect to this study, none of the anthropometric measures considered (somatotypes at age 7, 18 and one year prior to enrolment) were found to be associated with the eleven breast cancer susceptibility SNPs identified by Easton et al. [15] in the first GWAS of breast cancer risk.

Given the strength of the associations, and the ease of retrieval of information on childhood somatotypes retrospectively from pictures early in life, childhood body size is potentially useful for building breast cancer risk or prognosis prediction models. It appears counterintuitive that a large body size during childhood can reduce breast cancer risk or alter one's prognosis, because a large birth weight and a high adult BMI have been shown to otherwise elevate breast cancer risk. There remain unanswered questions on mechanisms driving this protective effect. Because body size and related hormonal exposures are modifiable risk factors, women might substantially decrease their risk of breast cancer, in particular the more aggressive ER-negative disease, by monitoring their nutrition and exogenous hormone intake at different points in life.

## 5.4 OTHER METHODOLOGICAL CONSTRAINTS

### 5.4.1 Study design

Statistics is neither computational conjuration nor mighty magic. Poor design can never be corrected by subtle analysis techniques. Below is a discussion on what was commendable about the studies, and a humble admission of our inadequacies in study design (with a heavy focus on the CAHRES study on which most of the analyses in this thesis was performed).

First of all, CAHRES is not a cohort study. Cohort studies are considered the gold standard of epidemiological studies. The advantages of a prospective cohort study is that i) the subjects are well-defined and selected randomly, and ii) since it follows every individual through time before the development of a certain disease or outcome, and collects data at regular intervals, recall bias is reduced. However, cohort studies are like heavy long-term fund investments: though rewarding, are expensive to set up and can take a long time to mature.

CAHRES is an example of the *opposite* of a cohort study – a case-control study. Contrary to a prospective cohort study, where exposure data is available but not outcome data, we already have data on the disease outcome – diagnosis of breast cancer and information on mammographic density. The information on exposure is collected retrospectively, through questionnaires or medical records. But memory could be vague, and imagination, strong. Hence, retrospective case-control studies are subjected to bias, which are discussed in further detail in the next section.

Life is dynamic, not static. Dietary and lifestyle habits can change, and so can physical measures. The only way to answer research questions pertaining to change is by collecting longitudinal data. While every effort has been taken to collect data from different time points in life, when Study III was performed, data on mammographic density was only available for the last screening prior to breast cancer diagnosis for cases, and the last available mammogram for controls.

### 5.4.2 Internal validity

*Bias and prejudice are attitudes to be kept in hand, not attitudes to be avoided.*
- Charles Curtis

Selection bias

*Control selection bias*. The beauty of the CAHRES dataset is that it is made up of cases and controls selected randomly from the general population. This is possible because of the well-curated cancer registries and unique personal numbers that every Swedish resident is given. This greatly reduces selection bias, which is defined as a statistical bias in which there is an error in choosing the individuals or groups to take part in a scientific study.

*Self-selection bias*. Some people don't give a hoot about politics. "If the election doesn't affect me, why vote?" It's the same mentality with case-control studies. Women diagnosed with breast cancer or women with relatives with breast cancer may be more motivated to participate in a study where many women may potentially benefit from the study results. On the other hand, healthy control women may be less aware of the disease, or feel less obliged to take time to fill out questionnaires or have their blood drawn (82% of invited women agreed to participate in study, of which 14% failed to return the mailed questionnaire). Behind the refusal to participate in the study could be differences in socioeconomic status (some women might be very much occupied by work, with little time left over to complete questionnaires etc.)

Apart from the possible blasé attitude of healthy women controls, eligible breast cancer cases could reject the invitation to participate in the study as well. Among all invited women with a diagnosis of breast cancer, 82% agreed to take part in this study. To reiterate what has been described in the Methods section, "non-participation was due to physcians' refusal (because of psychiatric disorder, anxiety or poor physical health in the patients), in 4% and patients' refusal (either to be approached at all or to return to questionnaire or failure in contacting the patient, in 12%". This non-participation can contribute to selection bias, as we are missing out women who had a more severe form of the disease. Hence, it would be unwise to generalize results from Studies I-IV to women with a more aggressive form of breast cancer.

Observation bias

*Recall bias.* While we can almost always remember if one of our loved ones ever had breast cancer, or readily measure our weight and height if we ever forget such information, it can sometimes be hard to recall what one had for dinner the night before, or how many cigarettes one smoked several decades ago. Included in the questionnaire that was sent out to all CAHRES participants were questions on the dietary and lifestyle habits ranging from childhood to one year prior to enrolment, and recalling such details could be tricky.

Recall bias could be differential, or non-differential. Differential recall bias occurs when either the cases or controls are more or less likely to remember and report prior exposures. Since the focus of this thesis is mainly on genetic variants, which are encoded in our DNA and determined by genotyping, neither cases nor controls could have contributed to differential recall bias.

Recalling the body size of a study subject when she was seven is a different matter. There could be a tendency for women to classify themselves a notch slimmer, but if all women tend to do that, the bias is non-differential. If breast cancer cases had prior knowledge that a certain somatotype predisposed one to breast cancer, they might differentially recall themselves to be closer to the "risky" somatotype. However, since somatotype is not commonly known as a breast cancer risk factor, the extent of such differential recall bias is likely minimal. Besides, the nine-level pictogram used in Study IV has been reported to be highly correlated with actual childhood measurements of height and weight in school.

Misclassification

*Measurement of estrogen receptor content.* Heterogeneity may arise from misclassification bias, especially since the ER and PR content of breast tumours were measured in seven different laboratories. This non-centralized testing of estrogen receptor expression could potentially lead to erroneous ER or PR classification for some patients, particularly for those with nuclear receptor expression levels close to or around the threshold. Different laboratories also used different methods for assessing nuclear receptor content. Exposure misclassification bias may result in an underestimation of the true associations.

*Reading of mammograms.* Misclassification of mammographic density may also arise from inter- and intra-individual variability. Given a mammogram, an untrained reader may produce different readings at two or more different occasions, since demarcation of breast or dense or non-dense areas are subjective. For each image in our study, a single trained observer (Louise Eriksson) set the appropriate gray-scale threshold levels defining the edge of the breast and distinguishing dense from nondense tissue. A set of mammograms were also used for “calibration” before each read, thus reducing the misclassification error in our study.

*Systematic genotyping error.* In the genome-wide studies, cases and controls from different populations were genotyped on different chips, and at different times, which could potentially introduce systematic errors when the data is combined and studied together. Errors in genotype determination can lead to bias in the estimation of genotype effects and reduces statistical power to detect true associations, thus increasing the required sample size. To minimize systematic genotyping error associated with each chip, we performed quality control on each genotyped set separately, scrutinized quantile-quantile plots of each population for systematic deviation from the diagonal $y=x$, applied correction for population stratification, and combined results via meta-analysis.

### 5.4.3 Statistical power and multiple testing

Epidemiological data was available for close to 3,000 cases and 3,000 controls in the CAHRES study. However, due to a constraint in resources, genotyping was only performed for half the total number of cases and controls in the candidate gene study (~1,500 cases, ~1,500 controls, Study III), and this number was further halved (~800 cases, ~800 controls) for the genome-wide studies (Studies I and II). For a hypothesis-generating stage I genome-wide association analysis, 800 non-familial cases do not boost a lot of statistical power, especially when hundreds of thousands of markers are being analyzed. To counter the lack of statistical power, we combined genotype data from other studies.

As for multiple testing, imagine the following fictitious scenario: A student took two successive tests on the same subject, one within an hour of the other. He barely passed the first one, yet scored a high distinction on the second attempt. The student was suspected of cheating and the grades were withheld by the examination committee. Feeling indignant, the student appealed to the head of the department, who has a background in statistics. The discussion was as such:

- Examination committee member

  “The chances of a student improving by such a great margin is 1 out of 100,000.”

- Head of department

  “Was there any other reason why you doubted the student’s performance?”

- Examination committee member

  “No.”

- Head of department

  “Did the invigilator find the student’s actions to be suspicious, or catch the student cheating during the exam?”

- Examination committee member

  “No.”

- Head of department

“How many students took the retest together with the student?”

- Examination committee member

“100,000.”

- Head of department

“Then, credit the student with the high distinction.”

In the genome-wide association studies described in this thesis, close to 300,000 SNPs were analyzed, of which upwards of 30,000 SNPs (10%) were found to be significantly associated at the 5% level. But because we looked at so many markers, 5% of the significantly associated SNPs could be so due to chance alone. Using the conventional threshold of 5% for chance events, we still do see an enrichment of significantly associated signals (10% - 5% = 5%). This excess of small p-values is good news, since it is indicative that we have some markers that are very likely to be truly associated with breast cancer within the 30,000.

For a result to stand out from the many markers tested in a genome-wide association study and worthy of costly further genotyping and validation in independent samples, it has to survive multiple testing. Using a conservative Bonferroni correction, for a marker to be considered significantly associated among 299,999 others that were found to be significant, it has to have a p-value that is smaller than 0.05/300,000, assuming an alpha of 0.05.

In a typical stage 2 or validation stage of genome-wide association studies, the most significantly associated hits (1,000 – 30,000 markers) are genotyped in an independent sample. This selection includes SNPs that do not survive testing, since true positives might be associated with a p-value anywhere ranging from 0.05 to 0.05/30,000. One of the main limitations of this study is that validation was only performed for the most significantly associated SNPs (less than ten for each of the GWAS). It is highly likely that our top hits were false positives; neither of the validation attempts was successful.

### 5.4.4 Wrapping up

In essence, what we have learnt is that:

#### Cancer is not just one mutation.

The principle of cooperativity runs deep in oncology. While we failed to add on to the list of breast cancer SNPs of fame, we hope to relay the message that genetic culprits of breast cancer may, in addition to single SNPs, be groups of SNPs forming a larger network.

#### Cancer is not just one phenotype.

In fact, no two cancers are alike. Cancers could be similar, in terms of subtypes like grade, ER or PR status, for example, but all cancer cells have a large number of abnormalities and it would be extremely rare for any two cancers to share the same mutation lineage, in the same tissue. Common genetic variants that are associated with the disease, when considered collectively, could help build an individual genetic portrait for personalized treatment.

The list of common breast cancer susceptibility loci currently known has been estimated to explain only 5% of the excess risk of breast cancer. Given the null findings of two GWAS mentioned in this abstract, and the likelihood that genetic effect sizes of any unfound variants are small, greater sample sizes and further studies are required to increase the variance explained for the disease. One could also attempt to increase the variance explained for the disease by looking for associations with strong predictors of breast cancer risk, such as mammographic density. Alternatively, as the data suggests, ER-negative breast cancer is a distinct breast cancer subtype that merits independent analyses, and could help to explain the missing heritability of breast cancer.

# 6 CONCLUSIONS

⇒ Pathway analysis of GWAS may help to prioritize the biological pathways most likely to be involved in the disease etiology. Further analysis of SNPs in pathways found to be associated with breast cancer risk in Study I may identify associations that would be difficult to detect through agnostic single SNP analyses. More effort focused in these aspects of oncology can potentially open up promising avenues for the understanding of breast cancer and its prevention.

⇒ ER-negative breast cancer is a distinct breast cancer subtype that merits independent analyses. Given the clinical importance of this phenotype and the likelihood that genetic effect sizes are small, greater sample sizes and further studies are required to understand the etiology of ER-negative breast cancers.

⇒ The way that SNPs are grouped and defined as pathways is important to the meaningful implementation of gene set enrichment analyses to SNP data.

⇒ Overall, there is no conclusive evidence that genetic variants in genes involved in the estrogen metabolic pathway are associated with mammographic density in postmenopausal women.

⇒ Given the limited understanding of the biology of mammographic density, the field should open up to hypothesis-free or discovery driven research. The trend in studying complex phenotypes is shifting more and more towards hypothesis-generating GWAS, where valuable new theories and understanding of the biology of a complex phenotype such as mammographic density may be originated.

⇒ Greater body size at age 7 is associated with a decreased risk of postmenopausal breast cancer, and the associated protective effect is stronger for the ER-negative breast cancer subtype than for the ER-positive subtype.

⇒ The bulk of knowledge on breast cancer risk is very often built upon static markers (e.g. a particular read of mammogram or measurement of BMI). Life course epidemiology or longitudinal studies have the potential to complement and improve existing data.

# 7 FINAL REMARKS AND FUTURE RESEARCH

The end of one journey is the beginning of another. While I hope to have contributed to the fight against breast cancer in my own way, the question remains: Where do we go from here?

### Functional relevance and the (Holy) GRAIL

Newly identified SNPs from hypothesis-generating GWAS require supporting validation and functional studies to make them shine. It is common practice to forward stellar hits with p-values less than $10^{-8}$ for validation in independent populations, but in doing so, we may miss out undiscovered gems of true association in the fuzzy region of small p-values that have not quite reach the level of genome-wide significance.

Apart from p-value rankings alone, functional relevance may come in as a potential mechanism to prioritize the hundreds of SNPs that are expected to achieve modest p-values of between $10^{-5}$ and $10^{-3}$ in a GWAS for follow-up. For example, a web-based text-mining tool developed by the Broad Institute, Gene Relationships Across Implicated Loci (GRAIL) [175], is able to suggest a degree of functional connectivity to each SNP or region. SNPs rated with high functional connectivity were shown to have a better success rate at being validated in independent data than SNPs rated with little or no functional connectivity.

### Invest in servers, software or technical expertise.

It is important to acknowledge the strategic role technologies can play in helping research communities overcome the challenges in unraveling the many mysteries of disease etiology that still remain. Technology moves fast. When GWAS data first came out a few years ago, ~100,000 markers seemed like quite a lot to handle (compared to genotyping by gel electrophoresis!). Data files could no longer be opened via the familiar Microsoft Excel, and files had to be sent to collaborators on DVDs. As the prices of genotyping fell with bigger and better technology, SNP chips that can produce more than a million genotypes for a fraction of the cost became commonplace. Data sets that used to be able to fit on one DVD have grown so big that it now takes a few DVDs for storage. One can only imagine shipping hard disks of data generated by next-generation sequencing to collaborators in the near future. While it is attractive to invest in genotyping/sequencing technologies, it is also prudent to set up infrastructure to house and process the upcoming data explosion. Otherwise, we might end up with a lot of data and no one to analyze it.

### Towards greater numbers for greater good.

It's not just one woman's fight. Nor is it that of any one student, one university, or one cancer research centre. As a large sample size is required to detect genetic or epidemiological risk factors that are only modestly associated with breast cancer risk, researchers all over the world are coming together to form gigantic, multi-centre collaborations in an effort to understand breast cancer. The next step would be to tap into the resources of breast cancer super-collaborations, be it the validation of our own hypothesis-generating GWAS, or the initiation of new projects.

For example, Collaborative Oncological Gene-Enivronment Study (COGS) [176] is a unification of established data sets of breast-, ovarian- and prostate cancers worldwide. Funded by the European Commission and 7th Framework Programme, the central focus of the project is to define individual risk of the three different cancers. Under the COGS initiative, we will have access to the largest breast cancer GWAS in the world, consisting of an estimated 120,000 individuals. Data collection for a special "cancer chip" called iCOGS which consists of ~200,000 SNPs contributed by breast cancer researchers all over the world is in place. Approximately 70,000 of the iCOGS SNPs are targeting the different areas of breast cancer inheritance. The results of this massive collaboration are expected to be ready by the end of spring, 2011, and holds great potential in unraveling the remaining mysteries of breast cancer.

The Karolinska Mammography (KARMA) study aims at generating a sufficiently large cohort of women to enable prediction of breast cancer risk. The cohort, consisting of only Swedish women at this stage, may also be used for randomized intervention studies where the effects of individualized screening, chemoprophylaxis or surgery would be assessed. A pilot project has just been initiated in the fall of 2010, with an estimated recruitment of 100,000 women by the year end of 2012, complete with information on repeated measurements of mammographic density, questionnaire data on risk factors, and blood samples drawn over several time points.

A similar effort is the ATHENA Breast Health Network [177], a revolutionary project which will initially involve 150,000 women in and around California. Participants will be screened for breast cancer and followed for decades through the five University of California medical centers. ATHENA is a University of California system-wide project supported by a $5.3 million University of California grant and a $4.8 million grant from the Safeway Foundation.

The ground-breaking projects above, and many more to come, are expected to generate a rich collection of data and knowledge that will shape breast cancer care in the way the renowned Framingham heart study changed the care of patients with heart disease.

# 8 AFTERWORD

## 8.1 IF I WERE A PROFESSOR…

As a student, it has been my incredible fortune and a whole load of luck that I get to attend many international conferences and meetings. While I hid in the shadows and listened to established giants of the field talk about cohort studies with obligatory interest, I couldn't help but feel detached - and disconnected.

The electronic mail is a boon to research. It costs almost nothing to communicate among scientists and scientists, and among scientists and data participants. It's fast, convenient and effective, and it shortens the time to collect data considerably.

Sure, there may be inherent bias in sample selection if we only approach the internet-savvy. But since we've already crossed the line, why should we stop at an email or online questionnaire?

I mentioned feeling disconnected. That's because I wasn't constantly checking emails, AND living my life on Facebook, AND laying out my everyday habits on clever iPhone apps that track my daily exercise, hormonal cycles, and what-nots. Send me an email inviting me to fill out a questionnaire to assess my breast cancer risk and ultimately do good for (wo)mankind? That could wait a day or two. But give me an app that's interactive and informative, and I might just get hooked!

The possibilities of introducing apps to data collection are endless. Difficult nutritional studies might actually be one step closer to being accurate and unbiased - if subjects could enter what they're having for meals just before they consume the food. Data may be stored and synced to a remote database and then processed directly. In return, the subjects gets processed data in the form of calorie counts, and rough pie charts on nutritional breakdown, analytical results on when they tend to "sin" when it comes to food binges, what they take too much or too little of. You can even throw in a neat reminder feature when meal times have passed and no entry has been recorded!

As a proof of concept, let's highlight the example of the wildly successful (in my own opinion) Nike+ app that, on top of archiving and analyzing all the miles you've accrued through running (Figure 8-1 and Figure 8-2), screams out your exercise routine on Facebook and Twitter (Figure 8-3). It even sneaks into your real life, as opposed to cyber life, through mini avatars (Figure 8-4) that prances out of screen-savers and guilt-trips you into breaking that sweat. The runner isn't the only one having all the fun and reaping all the personalized data - the Nike+ servers are a treasure cove of data waiting to be visualized, you just need someone to harvest it.

**Figure 8-1 Summary of the author's runs tracked on Nike+**
According to their website, "Add some personality, save every little detail and try to beat it next time."

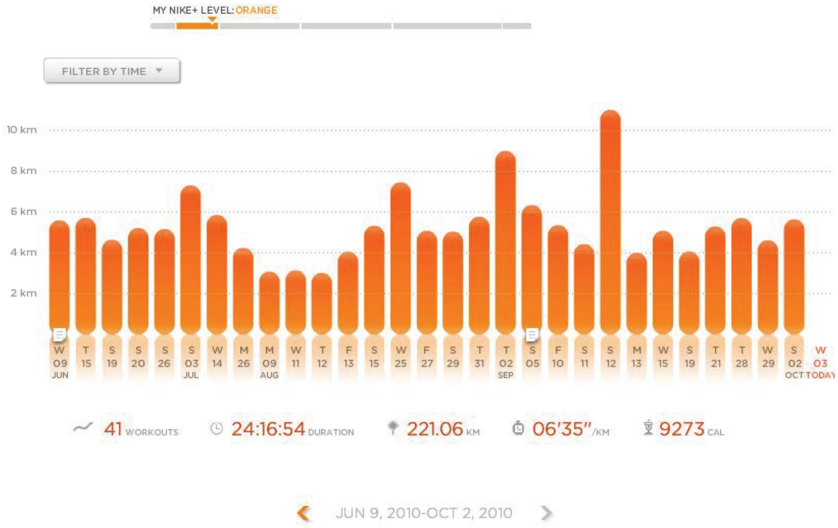
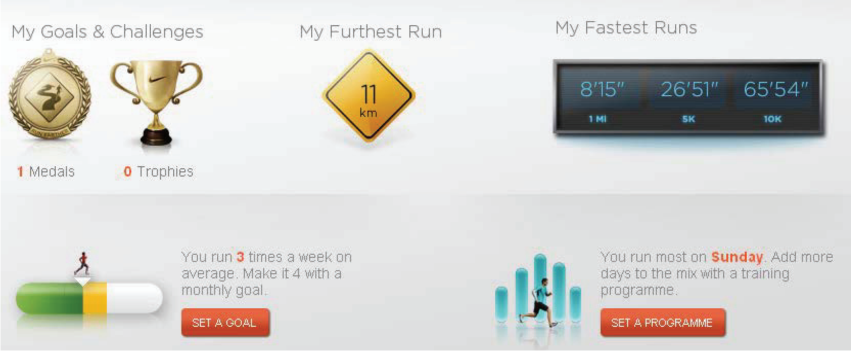


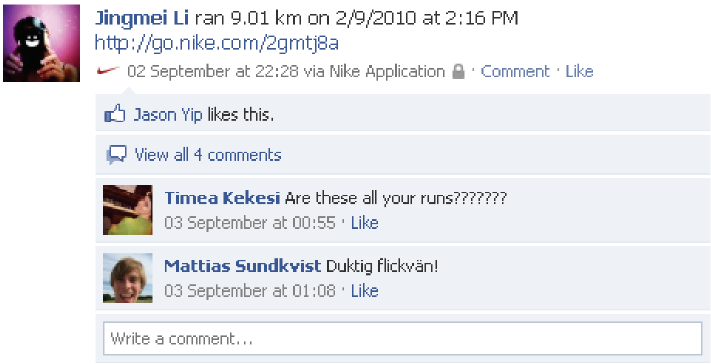
**Figure 8-2 Beat your best**
According to their website, "Set a goal, track your progress and find the motivation to become an even better runner."



You run 3 times a week on average. Make it 4 with a monthly goal.

You run most on Sunday. Add more days to the mix with a training programme.

**Figure 8-3 Screen shot of author's Facebook profile**
Half the fun is to boast about what you have achieved



Jingmei Li ran 9.01 km on 2/9/2010 at 2:16 PM
http://go.nike.com/2gmtj8a
02 September at 22:28 via Nike Application · Comment · Like

Jason Yip likes this.

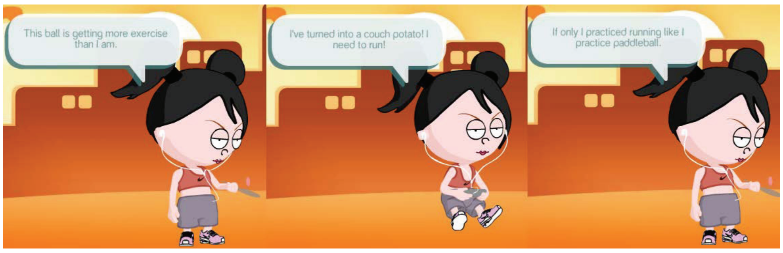
View all 4 comments

Timea Kekesi Are these all your runs???????
03 September at 00:55 · Like

Mattias Sundkvist Duktig flickvän!
03 September at 01:08 · Like

Write a comment...

**Figure 8-4 Different faces of the very temperamental mini avatar of the author's Nike+ profile**
For some, it gives all the motivation needed to put on those running shoes.


This ball is getting more exercise than I am.
I've turned into a couch potato! I need to run!
If only I practiced running like I practice paddleball.

Disclaimer: I am not doing an advertorial for Nike.

There are whimsical-sounding scientific pursuits that might just work in this exciting new age. Not belonging to Facebook, or carry a smart phone, is akin to not holding a driver's license of a government issued id. To not have an account is to say you don't have a personal phone number (gasp). If not already now, I won't bet against it happening in the future.

-Skickat från min iPhone

# 9 ACKNOWLEDGEMENTS

I am heartily thankful to my supervisors, Per Hall, Kamila Czene, Jianjun Liu and Keith Humphreys, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject.

I've also been helped by just about everyone I had a conversation with over the past few years. Every baby step I took was bigger than life, because I was standing on the shoulders of giants. So many people passed along their insights, or gave advice, or said or did something that gave me ideas or shed light on what I was doing. I am eternally grateful for the smiles and nods of acknowledgement in the hallway or pantry that light up the often dark winter days. Not to mention my family and friends who give me support and make me feel awesome, always.

I thank you. ☺

# 10 REFERENCES

1. **GLOBOCAN 2008, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10** [http://globocan.iarc.fr/]
2. Ferlay J, Bray F, Pisani P, Parkin DM: **GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide. IARC CancerBase No. 5. version 2.0**. In.: IARCPress, Lyon; 2004.
3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland**. *N Engl J Med* 2000, **343**(2):78-85.
4. Ghoussaini M, Pharoah PD: **Polygenic susceptibility to breast cancer: current state-of-the-art**. *Future Oncol* 2009, **5**(5):689-701.
5. **NORDCAN: Cancer Incidence, Mortality, Prevalence and Prediction in the Nordic Countries, Version 3.5.** [http://www.ancr.nu]
6. Singletary SE: **Rating the risk factors for breast cancer**. *Ann Surg* 2003, **237**(4):474-482.
7. Nelson NJ: **Migrant studies aid the search for factors linked to breast cancer risk**. *J Natl Cancer Inst* 2006, **98**(7):436-438.
8. Risch N: **The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches**. *Cancer Epidemiol Biomarkers Prev* 2001, **10**(7):733-741.
9. Locatelli I, Lichtenstein P, Yashin AI: **The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data**. *Twin Res* 2004, **7**(2):182-191.
10. Hunter K: **Host genetics influence tumour metastasis**. *Nat Rev Cancer* 2006, **6**(2):141-146.
11. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST *et al*: **Complement factor H polymorphism in age-related macular degeneration**. *Science* 2005, **308**(5720):385-389.
12. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S *et al*: **A genome-wide association study identifies novel risk loci for type 2 diabetes**. *Nature* 2007, **445**(7130):881-885.
13. Zhang FR, Huang W, Chen SM, Sun LD, Liu H, Li Y, Cui Y, Yan XX, Yang HT, Yang RD *et al*: **Genomewide association study of leprosy**. *N Engl J Med* 2009, **361**(27):2609-2618.
14. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proc Natl Acad Sci U S A* 2009, **106**(23):9362-9367.
15. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci**. *Nature* 2007, **447**(7148):1087-1093.
16. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A *et al*: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer**. *Nat Genet* 2007, **39**(7):870-874.

17. Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, Ponder BA: **Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer**. *PLoS Biol* 2008, **6**(5):e108.

18. Pharoah PD, Antoniou AC, Easton DF, Ponder BA: **Polygenes, risk prediction, and targeted prevention of breast cancer**. *N Engl J Med* 2008, **358**(26):2796-2803.

19. Reeves GK, Travis RC, Green J, Bull D, Tipper S, Baker K, Beral V, Peto R, Bell J, Zelenika D *et al*: **Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci**. *JAMA*, **304**(4):426-434.

20. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P *et al*: **Performance of common genetic variants in breast-cancer risk models**. *N Engl J Med*, **362**(11):986-993.

21. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ: **Projecting individualized probabilities of developing breast cancer for white females who are being examined annually**. *J Natl Cancer Inst* 1989, **81**(24):1879-1886.

22. Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P *et al*: **Genomewide association studies: history, rationale, and prospects for psychiatric disorders**. *Am J Psychiatry* 2009, **166**(5):540-556.

23. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.

24. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability and strategies for finding the underlying causes of complex disease**. *Nat Rev Genet*, **11**(6):446-450.

25. Gibson G: **Hints of hidden heritability in GWAS**. *Nat Genet*, **42**(7):558-560.

26. Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, Cleynen I, Colombel JF, de Rijk P, Dewit O *et al*: **Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease**. *Nat Genet*.

27. **Encyclopedia Britannica Academic Edition** [http://www.britannica.com/EBchecked/topic/1081410/genetic-heterogeneity]

28. van Spronsen FJ: **Phenylketonuria: a 21st century perspective**. *Nat Rev Endocrinol*, **6**(9):509-514.

29. Halperin E, Kimmel G, Shamir R: **Tag SNP selection in genotype data for maximizing SNP prediction accuracy**. *Bioinformatics* 2005, **21 Suppl 1**:i195-203.

30. 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**(11):1217-1223.

31. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Magi R *et al*: **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. *Nat Genet*, **42**(11):937-948.

32. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S *et al*: **Hundreds of variants clustered in genomic loci and biological pathways affect human height**. *Nature*, **467**(7317):832-838.

33. Hastings PJ, Lupski JR, Rosenberg SM, Ira G: **Mechanisms of change in gene copy number**. *Nat Rev Genet* 2009, **10**(8):551-564.
34. Eberharter A, Becker PB: **Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics**. *EMBO Rep* 2002, **3**(3):224-229.
35. Tamaru H, Selker EU: **A histone H3 methyltransferase controls DNA methylation in Neurospora crassa**. *Nature* 2001, **414**(6861):277-283.
36. Rose SP: **Commentary: heritability estimates--long past their sell-by date**. *Int J Epidemiol* 2006, **35**(3):525-527.
37. Maser JD, Akiskal HS: **Spectrum concepts in major mental disorders**. *Psychiatr Clin North Am* 2002, **25**(4):xi-xiii.
38. Yang YX, Wood NW, Latchman DS: **Molecular basis of Parkinson's disease**. *Neuroreport* 2009, **20**(2):150-156.
39. Anderson WF, Matsuno R: **Breast cancer heterogeneity: a mixture of at least two main types?** *J Natl Cancer Inst* 2006, **98**(14):948-951.
40. **Understanding Cancer Series: Estrogen Receptors/SERMS** [http://www.cancer.gov/cancertopics/understandingcancer/estrogenreceptors/]
41. Allred DC, Brown P, Medina D: **The origins of estrogen receptor alpha-positive and estrogen receptor alpha-negative human breast cancer**. *Breast Cancer Res* 2004, **6**(6):240-245.
42. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747-752.
43. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10869-10874.
44. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S *et al*: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8418-8423.
45. **Breast Cancer Facts and Figures 2009-2010** [http://www.cancer.org/acs/groups/content/@nho/documents/document/f861009final90809pdf.pdf]
46. Puliti D, Paci E: **The other side of technology: risk of overdiagnosis of breast cancer with mammography screening**. *Future Oncol* 2009, **5**(4):481-491.
47. Veronesi A, Serraino D: **Screening: is breast cancer overdiagnosed?** *Nat Rev Clin Oncol* 2009, **6**(12):682-683.
48. Welch HG, Black WC: **Overdiagnosis in cancer**. *J Natl Cancer Inst*, **102**(9):605-613.
49. Gram IT, Bremnes Y, Ursin G, Maskarinec G, Bjurstam N, Lund E: **Percentage density, Wolfe's and Tabar's mammographic patterns: agreement and association with risk factors for breast cancer**. *Breast Cancer Res* 2005, **7**(5):R854-861.
50. Boyd NF, Lockwood GA, Martin LJ, Knight JA, Byng JW, Yaffe MJ, Tritchler DL: **Mammographic densities and breast cancer risk**. *Breast Dis* 1998, **10**(3-4):113-126.

51. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ: **Mammographic densities and breast cancer risk**. *Cancer Epidemiol Biomarkers Prev* 1998, **7**(12):1133-1144.
52. Vachon CM, van Gils CH, Sellers TA, Ghosh K, Pruthi S, Brandt KR, Pankratz VS: **Mammographic density, breast cancer risk and risk prediction**. *Breast Cancer Res* 2007, **9**(6):217.
53. Stellman SD, Takezaki T, Wang L, Chen Y, Citron ML, Djordjevic MV, Harlap S, Muscat JE, Neugut AI, Wynder EL *et al*: **Smoking and lung cancer risk in American and Japanese men: an international case-control study**. *Cancer Epidemiol Biomarkers Prev* 2001, **10**(11):1193-1199.
54. Lagergren J, Bergstrom R, Lindgren A, Nyren O: **Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma**. *N Engl J Med* 1999, **340**(11):825-831.
55. Hadjisavvas A, Loizidou MA, Middleton N, Michael T, Papachristoforou R, Kakouri E, Daniel M, Papadopoulos P, Malas S, Marcou Y *et al*: **An investigation of breast cancer risk factors in Cyprus: a case control study**. *BMC Cancer*, **10**:447.
56. Wolfe JN: **Risk for breast cancer development determined by mammographic parenchymal pattern**. *Cancer* 1976, **37**(5):2486-2492.
57. Petroudi S, Kadir T, Brady M: **Automatic classification of mammographic parenchymal patterns: a statistical approach**. *Engineering in Medicine and Biology Society, 2003 Proceedings of the 25th Annual International Conference of the IEEE* 2003, **1**:798-801.
58. Gram IT, Funkhouser E, Tabar L: **The Tabar classification of mammographic parenchymal patterns**. *Eur J Radiol* 1997, **24**(2):131-136.
59. Wolfe JN, Saftlas AF, Salane M: **Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case-control study**. *AJR Am J Roentgenol* 1987, **148**(6):1087-1092.
60. Ursin G, Astrahan MA, Salane M, Parisky YR, Pearce JG, Daniels JR, Pike MC, Spicer DV: **The detection of changes in mammographic densities**. *Cancer Epidemiol Biomarkers Prev* 1998, **7**(1):43-47.
61. Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, Lockwood GA, Tritchler DL, Yaffe MJ: **Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study**. *J Natl Cancer Inst* 1995, **87**(9):670-675.
62. Garrido Estepa M, Ruiz-Perales F, Miranda J, Ascunce N, Gonzalez-Roman I, Sanchez-Contador C, Santamarina C, Moreo P, Vidal C, Peris M *et al*: **Evaluation of mammographic density patterns: reproducibility and concordance among scales**. *BMC Cancer*, **10**(1):485.
63. Devolli-Disha E, Manxhuka-Kerliu S, Ymeri H, Kutllovci A: **Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density**. *Bosn J Basic Med Sci* 2009, **9**(2):131-136.
64. Schrading S, Kuhl CK: **Mammographic, US, and MR imaging phenotypes of familial breast cancer**. *Radiology* 2008, **246**(1):58-70.
65. Boyd NF, Martin LJ, Bronskill M, Yaffe MJ, Duric N, Minkin S: **Breast tissue composition and susceptibility to breast cancer**. *J Natl Cancer Inst*, **102**(16):1224-1237.

66. Thompson DJ, Leach MO, Kwan-Lim G, Gayther SA, Ramus SJ, Warsi I, Lennard F, Khazen M, Bryant E, Reed S *et al*: **Assessing the usefulness of a novel MRI-based breast density estimation algorithm in a cohort of women at high genetic risk of breast cancer: the UK MARIBS study**. *Breast Cancer Res* 2009, **11**(6):R80.

67. Moore SG, Shenoy PJ, Fanucchi L, Tumeh JW, Flowers CR: **Cost-effectiveness of MRI compared to mammography for breast cancer screening in a high risk population**. *BMC Health Serv Res* 2009, **9**:9.

68. Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR, Giles GG, Tritchler D, Chiarelli A, Yaffe MJ *et al*: **Heritability of mammographic density, a risk factor for breast cancer**. *N Engl J Med* 2002, **347**(12):886-894.

69. Kelemen LE, Sellers TA, Vachon CM: **Can genes for mammographic density inform cancer aetiology?** *Nat Rev Cancer* 2008, **8**(10):812-823.

70. Lee E, Haiman CA, Ma H, Van Den Berg D, Bernstein L, Ursin G: **The role of established breast cancer susceptibility loci in mammographic density in young women**. *Cancer Epidemiol Biomarkers Prev* 2008, **17**(1):258-260.

71. Tamimi RM, Cox D, Kraft P, Colditz GA, Hankinson SE, Hunter DJ: **Breast cancer susceptibility loci and mammographic density**. *Breast Cancer Res* 2008, **10**(4):R66.

72. Woolcott CG, Maskarinec G, Haiman CA, Verheus M, Pagano IS, Le Marchand L, Henderson BE, Kolonel LN: **Association between breast cancer susceptibility loci and mammographic density: the Multiethnic Cohort**. *Breast Cancer Res* 2009, **11**(1):R10.

73. Odefrey F, Stone J, Gurrin LC, Byrnes GB, Apicella C, Dite GS, Cawson JN, Giles GG, Treloar SA, English DR *et al*: **Common genetic variants associated with breast cancer and mammographic density measures that predict disease**. *Cancer Res*, **70**(4):1449-1458.

74. **Science news. Genetic link between mammographic density and breast cancer.** [http://www.sciencedaily.com/releases/2010/02/100216113550.htm]

75. **The Melbourne Newsroom. Study reveals genetic link between mammographic density and breast cancer** [http://newsroom.melbourne.edu/studio/ep-71]

76. **Hospimedica. Genetic link found between mammographic density and breast cancer** [http://www.hospimedica.com/?Option=com_article&itemid=294728386&cat=women%27s%20health]

77. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS *et al*: **Genome-wide association study identifies five new breast cancer susceptibility loci**. *Nat Genet*, **42**(6):504-507.

78. Lord SJ, Mack WJ, Van Den Berg D, Pike MC, Ingles SA, Haiman CA, Wang W, Parisky YR, Hodis HN, Ursin G: **Polymorphisms in genes involved in estrogen and progesterone metabolism and mammographic density changes in women randomized to postmenopausal hormone therapy: results from a pilot study**. *Breast Cancer Res* 2005, **7**(3):R336-344.

79. Haiman CA, Bernstein L, Berg D, Ingles SA, Salane M, Ursin G: **Genetic determinants of mammographic density**. *Breast Cancer Res* 2002, **4**(3):R5.
80. Haiman CA, Hankinson SE, De Vivo I, Guillemette C, Ishibe N, Hunter DJ, Byrne C: **Polymorphisms in steroid hormone pathway genes and mammographic density**. *Breast Cancer Res Treat* 2003, **77**(1):27-36.
81. Biong M, Gram IT, Brill I, Johansen F, Solvang HK, Alnaes GI, Fagerheim T, Bremnes Y, Chanock SJ, Burdett L *et al*: **Genotypes and haplotypes in the insulin-like growth factors, their receptors and binding proteins in relation to plasma metabolic levels and mammographic density**. *BMC Med Genomics*, **3**:9.
82. Diorio C, Brisson J, Berube S, Pollak M: **Genetic polymorphisms involved in insulin-like growth factor (IGF) pathway in relation to mammographic breast density and IGF levels**. *Cancer Epidemiol Biomarkers Prev* 2008, **17**(4):880-888.
83. Taverne CW, Verheus M, McKay JD, Kaaks R, Canzian F, Grobbee DE, Peeters PH, van Gils CH: **Common genetic variation of insulin-like growth factor-binding protein 1 (IGFBP-1), IGFBP-3, and acid labile subunit in relation to serum IGF-I levels and mammographic density**. *Breast Cancer Res Treat*.
84. Diorio C, Sinotte M, Brisson J, Berube S, Pollak M: **Vitamin D pathway polymorphisms in relation to mammographic breast density**. *Cancer Epidemiol Biomarkers Prev* 2008, **17**(9):2505-2508.
85. Kelemen LE, Pankratz VS, Sellers TA, Brandt KR, Wang A, Janney C, Fredericksen ZS, Cerhan JR, Vachon CM: **Age-specific trends in mammographic density: the Minnesota Breast Cancer Family Study**. *Am J Epidemiol* 2008, **167**(9):1027-1036.
86. Nielsen M, Pettersen PC, Alexandersen P, Karemore G, Raundahl J, Loog M, Christiansen C: **Breast density changes associated with postmenopausal hormone therapy: post hoc radiologist- and computer-based analyses**. *Menopause*, **17**(4):772-778.
87. van Duijnhoven FJ, Peeters PH, Warren RM, Bingham SA, van Noord PA, Monninkhof EM, Grobbee DE, van Gils CH: **Postmenopausal hormone therapy and changes in mammographic density**. *J Clin Oncol* 2007, **25**(11):1323-1328.
88. van der Maas PJ, de Koning HJ, van Ineveld BM, van Oortmarssen GJ, Habbema JD, Lubbe KT, Geerts AT, Collette HJ, Verbeek AL, Hendriks JH *et al*: **The cost-effectiveness of breast cancer screening**. *Int J Cancer* 1989, **43**(6):1055-1060.
89. Knox EG: **Evaluation of a proposed breast cancer screening regimen**. *BMJ* 1988, **297**(6649):650-654.
90. **The Human Genome Project** [http://www.genome.gov/11006943]
91. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, Berdasco M, Fraga MF, O'Hanlon TP, Rider LG *et al*: **Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus**. *Genome Res*, **20**(2):170-179.
92. Levenson JM, Sweatt JD: **Epigenetic mechanisms in memory formation**. *Nat Rev Neurosci* 2005, **6**(2):108-118.
93. Avise JC, Mank JE: **Evolutionary perspectives on hermaphroditism in fishes**. *Sex Dev* 2009, **3**(2-3):152-163.

94. Desjardins JK, Fernald RD: **Fish sex: why so diverse?** *Curr Opin Neurobiol* 2009, **19**(6):648-653.

95. Richards EJ: **Inherited epigenetic variation--revisiting soft inheritance**. *Nat Rev Genet* 2006, **7**(5):395-401.

96. Sollars V, Lu X, Xiao L, Wang X, Garfinkel MD, Ruden DM: **Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution**. *Nat Genet* 2003, **33**(1):70-74.

97. Jablonka E, Raz G: **Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution**. *Q Rev Biol* 2009, **84**(2):131-176.

98. Bygren LO, Kaati G, Edvinsson S: **Longevity determined by paternal ancestors' nutrition during their slow growth period**. *Acta Biotheor* 2001, **49**(1):53-59.

99. Kaati G, Bygren LO, Edvinsson S: **Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period**. *Eur J Hum Genet* 2002, **10**(11):682-688.

100. Pembrey ME, Bygren LO, Kaati G, Edvinsson S, Northstone K, Sjostrom M, Golding J: **Sex-specific, male-line transgenerational responses in humans**. *Eur J Hum Genet* 2006, **14**(2):159-166.

101. Thinggaard M, Jacobsen R, Jeune B, Martinussen T, Christensen K: **Is the relationship between BMI and mortality increasingly U-shaped with advancing age? A 10-year follow-up of persons aged 70-95 years**. *J Gerontol A Biol Sci Med Sci*, **65**(5):526-531.

102. Kloner RA, Rezkalla SH: **To drink or not to drink? That is the question**. *Circulation* 2007, **116**(11):1306-1317.

103. Uauy R, Solomons N: **Diet, nutrition, and the life-course approach to cancer prevention**. *J Nutr* 2005, **135**(12 Suppl):2934S-2945S.

104. Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, Zoeller RT, Gore AC: **Endocrine-disrupting chemicals: an Endocrine Society scientific statement**. *Endocr Rev* 2009, **30**(4):293-342.

105. Ronckers CM, Erdmann CA, Land CE: **Radiation and breast cancer: a review of current evidence**. *Breast Cancer Res* 2005, **7**(1):21-32.

106. Land CE: **Studies of cancer and radiation dose among atomic bomb survivors. The example of breast cancer**. *JAMA* 1995, **274**(5):402-407.

107. Lambe M, Hsieh C, Trichopoulos D, Ekbom A, Pavia M, Adami HO: **Transient increase in the risk of breast cancer after giving birth**. *N Engl J Med* 1994, **331**(1):5-9.

108. Ahlgren M, Melbye M, Wohlfahrt J, Sorensen TI: **Growth patterns and the risk of breast cancer in women**. *N Engl J Med* 2004, **351**(16):1619-1626.

109. Xu X, Dailey AB, Peoples-Sheps M, Talbott EO, Li N, Roth J: **Birth weight as a risk factor for breast cancer: a meta-analysis of 18 epidemiological studies**. *J Womens Health (Larchmt)* 2009, **18**(8):1169-1178.

110. McCormack VA, dos Santos Silva I, De Stavola BL, Mohsen R, Leon DA, Lithell HO: **Fetal growth and subsequent risk of breast cancer: results from long term follow up of Swedish cohort**. *BMJ* 2003, **326**(7383):248.

111. Eden JA: **Breast cancer, stem cells and sex hormones: part 1. The impact of fetal life and infancy**. *Maturitas*, **67**(2):117-120.

112. Tamimi RM, Lagiou P, Czene K, Liu J, Ekbom A, Hsieh CC, Adami HO, Trichopoulos D, Hall P: **Birth weight, breast cancer susceptibility loci, and breast cancer risk**. *Cancer Causes Control*, **21**(5):689-696.

113. Berkey CS, Frazier AL, Gardner JD, Colditz GA: **Adolescence and breast carcinoma risk**. *Cancer* 1999, **85**(11):2400-2409.

114. Baer HJ, Tworoger SS, Hankinson SE, Willett WC: **Body fatness at young ages and risk of breast cancer throughout life**. *Am J Epidemiol*, **171**(11):1183-1194.

115. Hilakivi-Clarke L, Forsen T, Eriksson JG, Luoto R, Tuomilehto J, Osmond C, Barker DJ: **Tallness and overweight during childhood have opposing effects on breast cancer risk**. *Br J Cancer* 2001, **85**(11):1680-1684.

116. Kawai M, Minami Y, Kuriyama S, Kakizaki M, Kakugawa Y, Nishino Y, Ishida T, Fukao A, Tsuji I, Ohuchi N: **Adiposity, adult weight change and breast cancer risk in postmenopausal Japanese women: the Miyagi Cohort Study**. *Br J Cancer*.

117. Mathers JC: **Early nutrition: impact on epigenetics**. *Forum Nutr* 2007, **60**:42-48.

118. Hilakivi-Clarke L: **Nutritional modulation of terminal end buds: its relevance to breast cancer prevention**. *Curr Cancer Drug Targets* 2007, **7**(5):465-474.

119. De Assis S, Hilakivi-Clarke L: **Timing of dietary estrogenic exposures and breast cancer risk**. *Ann N Y Acad Sci* 2006, **1089**:14-35.

120. Parr CL, Batty GD, Lam TH, Barzi F, Fang X, Ho SC, Jee SH, Ansary-Moghaddam A, Jamrozik K, Ueshima H *et al*: **Body-mass index and cancer mortality in the Asia-Pacific Cohort Studies Collaboration: pooled analyses of 424,519 participants**. *Lancet Oncol*, **11**(8):741-752.

121. Grossmann ME, Ray A, Nkhata KJ, Malakhov DA, Rogozina OP, Dogan S, Cleary MP: **Obesity and breast cancer: status of leptin and adiponectin in pathological processes**. *Cancer Metastasis Rev*.

122. Cleary MP, Grossmann ME: **Minireview: Obesity and breast cancer: the estrogen connection**. *Endocrinology* 2009, **150**(6):2537-2542.

123. Borgquist S, Jirstrom K, Anagnostaki L, Manjer J, Landberg G: **Anthropometric factors in relation to different tumor biological subgroups of postmenopausal breast cancer**. *Int J Cancer* 2009, **124**(2):402-411.

124. Simpson ER: **Sources of estrogen and their importance**. *J Steroid Biochem Mol Biol* 2003, **86**(3-5):225-230.

125. Strobl JS, Wonderlin WF, Flynn DC: **Mitogenic signal transduction in human breast cancer cells**. *Gen Pharmacol* 1995, **26**(8):1643-1649.

126. Saxena T, Lee E, Henderson KD, Clarke CA, West D, Marshall SF, Deapen D, Bernstein L, Ursin G: **Menopausal hormone therapy and subsequent risk of specific invasive breast cancer subtypes in the California Teachers Study**. *Cancer Epidemiol Biomarkers Prev*, **19**(9):2366-2378.

127. Bardia A, Vachon CM, Olson JE, Vierkant RA, Wang AH, Hartmann LC, Sellers TA, Cerhan JR: **Relative weight at age 12 and risk of postmenopausal breast cancer**. *Cancer Epidemiol Biomarkers Prev* 2008, **17**(2):374-378.

128. Iwasaki M, Otani T, Inoue M, Sasazuki S, Tsugane S: **Body size and risk for breast cancer in relation to estrogen and progesterone receptor status in Japan**. *Ann Epidemiol* 2007, **17**(4):304-312.

129. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.

130. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR *et al*: **TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study**. *N Engl J Med* 2007, **357**(12):1199-1209.

131. Bengtsson C, Berglund A, Serra ML, Nise L, Nordmark B, Klareskog L, Alfredsson L, Eira Study G: **Non-participation in EIRA: a population-based case-control study of rheumatoid arthritis**. *Scand J Rheumatol*.

132. Colditz GA, Hankinson SE: **The Nurses' Health Study: lifestyle and health among women**. *Nat Rev Cancer* 2005, **5**(5):388-396.

133. Hankinson SE, Willett WC, Manson JE, Colditz GA, Hunter DJ, Spiegelman D, Barbieri RL, Speizer FE: **Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women**. *J Natl Cancer Inst* 1998, **90**(17):1292-1299.

134. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S *et al*: **A common coding variant in CASP8 is associated with breast cancer risk**. *Nat Genet* 2007, **39**(3):352-358.

135. Rosenberg LU, Einarsdottir K, Friman EI, Wedren S, Dickman PW, Hall P, Magnusson C: **Risk factors for hormone receptor-defined breast cancer in postmenopausal women**. *Cancer Epidemiol Biomarkers Prev* 2006, **15**(12):2482-2488.

136. Orgeas CC, Hall P, Rosenberg LU, Czene K: **The influence of menstrual risk factors on tumor characteristics and survival in postmenopausal breast cancer**. *Breast Cancer Res* 2008, **10**(6):R107.

137. Brouillet JP, Dujardin MA, Chalbos D, Rey JM, Grenier J, Lamy PJ, Maudelonde T, Pujol P: **Analysis of the potential contribution of estrogen receptor (ER) beta in ER cytosolic assay of breast cancer**. *Int J Cancer* 2001, **95**(4):205-208.

138. Tamimi RM, Eriksson L, Lagiou P, Czene K, Ekbom A, Hsieh CC, Adami HO, Trichopoulos D, Hall P: **Birth weight and mammographic density among postmenopausal women in Sweden**. *Int J Cancer*, **126**(4):985-991.

139. Byng JW, Boyd NF, Little L, Lockwood G, Fishell E, Jong RA, Yaffe MJ: **Symmetry of projection in the quantitative analysis of mammographic images**. *Eur J Cancer Prev* 1996, **5**(5):319-327.

140. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A: **The SNP ratio test: pathway analysis of genome-wide association datasets**. *Bioinformatics* 2009, **25**(20):2762-2763.

141. Tyrer J, Pharoah PD, Easton DF: **The admixture maximum likelihood test: a novel experiment-wise test of association between disease and multiple SNPs**. *Genet Epidemiol* 2006, **30**(7):636-643.

142. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P: **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder**. *Nature* 2009, **460**(7256):748-752.

143. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38**(8):904-909.

144. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ: **SCAN: SNP and copy number annotation**. *Bioinformatics*, **26**(2):259-262.

145. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS: **PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W399-405.

146. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

147. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap**. *Bioinformatics* 2008, **24**(24):2938-2939.

148. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**(2):263-265.

149. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ: **LocusZoom: regional visualization of genome-wide association scan results**. *Bioinformatics*, **26**(18):2336-2337.

150. **Edraw Mindmap** [http://www.edrawsoft.com/freemind.php]

151. R Development Core Team: **R: A Language and Environment for Statistical Computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2007.

152. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.

153. Storey JD, Tibshirani R: **Statistical methods for identifying differentially expressed genes in DNA microarrays**. *Methods Mol Biol* 2003, **224**:149-157.

154. Gauderman W, Morrison J: **QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe**. In.; 2006.

155. Li J, Humphreys K, Heikkinen T, Aittomaki K, Blomqvist C, Pharoah PD, Dunning AM, Ahmed S, Hooning MJ, Martens JW *et al*: **A combined analysis of genome-wide association studies in breast cancer**. *Breast Cancer Res Treat*.

156. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, Kraft P, Hunter DJ, Chanock SJ, Rosenberg PS *et al*: **Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade**. *Cancer Res*, **70**(11):4453-4459.

157. Li J, Eriksson L, Humphreys K, Czene K, Liu J, Tamimi RM, Lindstrom S, Hunter DJ, Vachon CM, Couch FJ *et al*: **Genetic variation in the estrogen metabolic pathway and mammographic density as an intermediate phenotype of breast cancer**. *Breast Cancer Res*, **12**(2):R19.

158. Li J, Humphreys K, Eriksson L, Czene K, Liu J, Hall P: **Effects of childhood body size on breast cancer tumour characteristics**. *Breast Cancer Res*, **12**(2):R23.

159. Low YL, Li Y, Humphreys K, Thalamuthu A, Darabi H, Wedren S, Bonnard C, Czene K, Iles MM, Heikkinen T *et al*: **Multi-variant pathway association analysis reveals the importance of genetic determinants of estrogen metabolism in breast and endometrial cancer susceptibility**. *PLoS Genet*, **6**:e1001012.

160. Hao K, Chudin E, Greenawalt D, Schadt EE: **Magnitude of stratification in human populations and impacts on genome wide association studies**. *PLoS One*, **5**(1):e8695.

161. Wang K, Bucan M, Grant SF, Schellenberg G, Hakonarson H: **Strategies for genetic studies of complex diseases**. *Cell*, **142**(3):351-353; author reply 353-355.

162. Dai JY, Ruczinski I, LeBlanc M, Kooperberg C: **Imputation methods to improve inference in SNP association studies**. *Genet Epidemiol* 2006, **30**(8):690-702.

163. **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.

164. Marchini J, Howie B: **Genotype imputation for genome-wide association studies**. *Nat Rev Genet*, **11**(7):499-511.

165. Huang L, Wang C, Rosenberg NA: **The relationship between imputation error and statistical power in genetic association studies in diverse populations**. *Am J Hum Genet* 2009, **85**(5):692-698.

166. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, Edenberg HJ, Kramer JR, A MG, Bierut LJ *et al*: **A new statistic to evaluate imputation reliability**. *PLoS One*, **5**(3):e9697.

167. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP: **Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms**. *Am J Hum Genet* 2008, **83**(1):112-119.

168. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF: **Practical aspects of imputation-driven meta-analysis of genome-wide association studies**. *Hum Mol Genet* 2008, **17**(R2):R122-128.

169. Torkamani A, Topol EJ, Schork NJ: **Pathway analysis of seven common diseases assessed by genome-wide association**. *Genomics* 2008, **92**(5):265-272.

170. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS results: A review of statistical methods and recommendations for their application**. *Am J Hum Genet*, **86**(1):6-22.

171. Schwender H, Ruczinski I, Ickstadt K: **Testing SNPs and sets of SNPs for importance in association studies**. *Biostatistics*.

172. **MetaCore™** [http://www.genego.com/metacore.php]

173. **BioCarta** [www.biocarta.com/]

174. **National Cancer Institute (NCI) Pathway Interaction Database** [http://pid.nci.nih.gov/]

175. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ: **Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions**. *PLoS Genet* 2009, **5**(6):e1000534.

176. **Collaborative Oncological Gene-environment Study** [www.cogseu.org/]

177. **ATHENA Breast Health Network** [www.athenacarenetwork.org/ ]

# I

Breast Cancer Res Treat
DOI 10.1007/s10549-010-1172-9

EPIDEMIOLOGY

# A combined analysis of genome-wide association studies in breast cancer

Jingmei Li · Keith Humphreys · Tuomas Heikkinen · Kristiina Aittomäki · Carl Blomqvist · Paul D. P. Pharoah · Alison M. Dunning · Shahana Ahmed · Maartje J. Hooning · John W. M. Martens · Ans M. W. van den Ouweland · Lars Alfredsson · Aarno Palotie · Leena Peltonen-Palotie · Astrid Irwanto · Hui Qi Low · Garrett H. K. Teoh · Anbupalam Thalamuthu · Douglas F. Easton · Heli Nevanlinna · Jianjun Liu · Kamila Czene · Per Hall

Received: 8 April 2010 / Accepted: 9 September 2010

**Abstract** In an attempt to identify common disease susceptibility alleles for breast cancer, we performed a combined analysis of three genome-wide association studies (GWAS), involving 2,702 women of European ancestry with invasive breast cancer and 5,726 controls. Tests for association were performed for 285,984 SNPs. Evidence for association with SNPs in genes in specific pathways was assessed using a permutation-based approach. We confirmed associations with loci reported by previous GWAS on 1p11.2, 2q35, 3p, 5p12, 8q24, 10q23.13, 14q24.1 and 16q. Six SNPs with the strongest signals of association with breast cancer, and which have not been reported previously, were typed in two further studies; however, none of the associations could be confirmed. Suggestive evidence for an excess of associations was found for genes involved in the regulation of actin cytoskeleton, glycan degradation, alpha-linolenic acid metabolism, circadian rhythm, hematopoietic cell lineage and drug metabolism. Androgen and oestrogen metabolism, a pathway previously found to be associated with the development of postmenopausal breast cancer, was marginally significant ($P = 0.051$ [unadjusted]). These

**Electronic supplementary material** The online version of this article (doi:10.1007/s10549-010-1172-9) contains supplementary material, which is available to authorized users.

J. Li · K. Humphreys · K. Czene · P. Hall (✉)
Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, P.O. Box 281, 17177 Stockholm, Sweden
e-mail: Per.Hall@ki.se

J. Li · A. Irwanto · H. Q. Low · G. H. K. Teoh ·
A. Thalamuthu · J. Liu (✉)
Human Genetics, Genome Institute of Singapore, 60 Biopolis
Street, #02-01 Genome, Singapore 138672, Singapore
e-mail: liuj3@gis.a-star.edu.sg

T. Heikkinen · H. Nevanlinna
Department of Obstetrics and Gynecology, Helsinki University
Central Hospital, P.O. Box 700, 00029 HUS Helsinki, Finland

K. Aittomäki
Department of Clinical Genetics, Helsinki University Central
Hospital, P.O. Box 140, 00029 HUS Helsinki, Finland

C. Blomqvist
Department of Oncology, Helsinki University Central Hospital,
P.O. Box 180, 00029 HUS Helsinki, Finland

P. D. P. Pharoah · D. F. Easton
Department of Public Health and Primary Care,
University of Cambridge, Cambridge CB1 8RN, UK

P. D. P. Pharoah · A. M. Dunning · S. Ahmed · D. F. Easton
Department of Oncology, University of Cambridge,
Cambridge CB1 8RN, UK

M. J. Hooning · J. W. M. Martens
Department of Medical Oncology, Rotterdam Family Cancer
Clinic, Erasmus University Medical Center, Rotterdam,
Netherlands

A. M. W. van den Ouweland
Department of Clinical Genetics, Rotterdam Family Cancer
Clinic, Erasmus University Medical Center, Rotterdam,
Netherlands

L. Alfredsson
Institute of Environmental Medicine, Karolinska Institutet,
P.O. Box 281, Stockholm 17177, Sweden

A. Palotie · L. Peltonen-Palotie
Institute for Molecular Medicine Finland, FIMM,
University of Helsinki, P.O. Box 20, 00014 Helsinki, Finland

A. Palotie · L. Peltonen-Palotie
Public Health Genomics Unit, National Institute for Health
and Welfare, P.O. Box 30, 00271 Helsinki, Finland

Published online: 26 September 2010

results suggest that further analysis of SNPs in these pathways may identify associations that would be difficult to detect through agnostic single SNP analyses. More effort focused in these aspects of oncology can potentially open up promising avenues for the understanding of breast cancer and its prevention.

**Keywords** Breast neoplasms · Genetic association studies · Genetic epidemiology · Genetic susceptibility · Genetic predisposition to disease/genetics · Case–control studies

**Abbreviations**

| | |
|---|---|
| CGEMS | Cancer Genetic Markers of Susceptibility |
| CI | Confidence interval |
| EIRA | Epidemiological Investigation of Rheumatoid Arthritis |
| FPRP | False positive report probability |
| GWAS | Genome-wide association study |
| OR | Odds ratio |
| RBCS | Rotterdam Breast Cancer Study |
| SEARCH | Study of Epidemiology and Risk factors in Cancer Heredity |
| SNP | Single nucleotide polymorphism |
| $\lambda_{GC}$ | Genomic inflation factor $\lambda$ |

## Introduction

Genome-wide association studies (GWAS) interrogating up to half a million markers have indentified low-penetrance, common genetic variants in 12 genomic regions that predispose for breast cancer [1–7]. The majority of these studies have been conducted with impressive collaborative efforts on pooled genotype data of more than 30,000 individuals in the validation stage. Despite these efforts, the identified loci can only explain about 5% of the excess familial risk of breast cancer [8], suggesting that many common and rare variants of similar or smaller effect on the disease remains to be identified. Under the assumption that the strongest association signals have already been found, and given the stringent significance thresholds required for GWAS, each new GWAS has limited power to identify new loci. Thus, extremely large GWAS, or combined analysis of GWAS, will be required to identify further loci by this approach.

An alternative approach to identify new loci is to subject GWAS to pathway-based analysis [9]. If a specific pathway is relevant to disease susceptibility, one might expect association signals to be overrepresented among single nucleotide polymorphisms (SNPs) in genes in the pathway. Assessment of the overall significance of SNPs in a given pathway circumvents some of the multiple testing problems, and offers the potential to identify loci that would be too weak to find through single SNP analysis. In addition, it may provide additional insights into the mechanisms underlying disease susceptibility [10].

In this study, we conducted a GWAS on 2,702 women of European ancestry with invasive breast cancer and 5,726 controls. We also examined the significance of pre-defined, biologically meaningful pathways on breast cancer risk using these data.

## Methods summary

Full methods accompany this manuscript in Supplementary File 1.

### Study populations

Table 1 summarises the origins, and numbers, of cases and controls used in this study. Subjects were drawn from three independent populations from Sweden, Finland and the National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMS) [3] initiative.

The Swedish sample set included 803 breast cancer cases and 764 controls which were drawn from a parent population-based cascontrol study of postmenopausal breast cancer which has been described elsewhere [11, 12]. An additional 659 cancer-free Swedish controls aged between 18 and 70 years were obtained from the Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study [13], primarily to improve statistical power.

The Finnish breast cancer study population consists of two series of unselected breast cancer patients and additional familial cases ascertained at the Helsinki University Central Hospital. We genotyped a total of 782 breast cancer cases from this study. Of these women, 212 were premenopausal, 359 were postmenopausal and 211 were missing menopausal status. Population control data was obtained from the Finnish Genome center on 3,170 healthy population controls described in [14–17].

Genotype data was also obtained for a total of 1,145 postmenopausal women of European ancestry with invasive breast cancer from the CGEMS initiative, along with

A. Palotie · L. Peltonen-Palotie
Wellcome Trust Sanger Institute, Hinxton,
Cambridge CB10 1SA, UK

A. Palotie · L. Peltonen-Palotie
Program in Medical and Population Genetics,
Broad Institute of Harvard and Massachusetts Institute
of Technology, Cambridge, MA 02142, USA

**Table 1** Summary of samples and genotyping platforms used in the discovery stage

| Study | Type | No. of samples (after quality control) | Genotyping platform |
|---|---|---|---|
| Swedish | Cases | 803 (797) | HumanHap300 supplemented by HumanHap240S |
| | Controls | 764 (764) | HumanHap550 |
| | Additional controls from EIRA study | 659 (650) | HumanHap300 |
| Finnish | Cases | 782 (760) | HumanHap550 |
| | Controls | 3170 (3170) | HumanHap370Duo |
| CGEMS | Cases | 1145 (1145) | HumanHap550 |
| | Controls | 1142 (1142) | HumanHap550 |

1,142 matched controls nested within the prospective Nurses' Health Study cohort [3].

For all populations, blood samples were obtained from individuals according to protocols and informed-consent procedures approved by institutional review boards.

### Genotyping and quality control filters

Genotyping for all samples was performed according to the Illumina Infinium 2 assay manual (Illumina, San Diego), as described previously [18]. The genotyping platforms used for this study are listed in Table 1. Apart from the 3,170 Finnish controls which were genotyped on the HumanHap370Duo assay as described previously [14, 16], genotyping for all other Finnish and Swedish samples was performed at the Genome Institute of Singapore. All Swedish cases were genotyped on the HumanHap300 platform and supplemented by the HumanHap240S platform. Swedish controls were genotyped on the HumanHap550 platform. The EIRA scan included genotypes of 317,503 SNPs from the HumanHap300 arrays. Details on the genotyping of CGEMS subjects using the HumanHap550 platform have been reported elsewhere [3].

Each dataset was filtered to remove individuals with >10% missing genotypes, and SNPs with >10% missing data, or minor allele frequency (MAF) <0.03, or not in Hardy–Weinberg equilibrium (HWE) ($P < 0.05$/number of SNPs after quality control) and individual samples with evidence of possible DNA contamination, common ancestry or cryptic family relationships. Quality control was carried out using the software Plink [19]. To account for population outliers and correct for differential ancestry between cases and controls that may exist in the dataset after familial outlier removal, a principal component (PC) analysis was conducted using the EIGENSTRAT software [20]. A total of 2,702 cases and 5,726 controls passed the quality control for samples. The 285,984 SNPs that passed quality control filters in all sample sets were merged into a single file for analysis. The merged dataset was subjected to the same quality control checks as carried out for the individual datasets. As an additional quality control check, genotype cluster plots of the top SNPs (lowest $P$ values) from the discovery stage were inspected manually using Illumina Beadstudio software to confirm the genotype calling.

### Statistical analysis

Logistic regression models with genotype coded 0, 1, 2 and treated as a continuous covariate (one at a time), were fitted for each SNP that passed quality control. An additive genetic effect on the logit scale was assumed to characterise the associations. Other genetic models, namely, dominant and recessive models, were also explored. Eigenvalues of PCs were included as covariates. Separate analyses were performed for the Swedish, Finnish and CGEMS datasets, together with a combined analysis. In the combined analysis, the final model included as covariates the SNP genotype, an indicator variable specifying country (Sweden, Finland and USA), and interaction effects of eigenvalues of PCs × country specified in such a way that country-specific PCs were implemented for the relevant subjects. Quantile–quantile plots were used to check for systematic genotyping error or bias due to unaccounted underlying population substructure. Manhattan plots were generated to summarise the −log transformed $P$ values of all SNPs examined. Association results corrected for genomic control and multiple testing are available as online-only supplementary material. Pair-wise linkage disequilibrium (LD) was evaluated for the top SNPs that were observed to be located in the same chromosomal region using Plink [19]. Tests of homogeneity of odds ratios across strata (Cochrane's $Q$ statistic and $I^2$ hetergeneity index) were conducted. The six most strongly associated SNPs in the combined analysis which had effects in the same direction for all three studies (Swedish, Finnish and CGEMS), and for which associations have not been described previously, were forwarded for validation in independent studies. All SNP chromosomal positions were based on NCBI Build 36.

Pathway analysis of the GWAS dataset was conducted using the SNP ratio test (SRT) [21]. The same logistic regression models which were applied to the real dataset were applied to 1000 datasets in which phenotypes were permuted, in order to obtain *P* value estimates. SRT was used to investigate the associations with breast cancer for 212 pathways and their genes (~4,700) taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (05/12/08) [22]. SNP to gene mappings were obtained by parsing the dbSNP table b129_SNPContigLocusId_36_3.bcp. This includes SNPs <2 kb 5′ and <0.5 kb 3′ of a gene. We applied the false discovery rate (FDR) *Q* value multiple testing correction [23, 24] to all empirical *P* value outputs from SRT to account for the large number of pathway definitions tested against the data. The default smoother method in the QVALUE software was applied.

PLINK (v1.06) [19], SNP Ratio Test [21], R (v2.8.0) [25], QVALUE [23], HaploView [26], PolySearch [27] and SCAN [28] were used for data management, quality control, statistical analyses, graphics, text-mining and SNP annotation purposes.

### Validation

Six SNPs with the strongest signals of association with breast cancer computed under the assumption of an additive model, and which have not been reported previously, were typed in two further studies: the Study of Epidemiology and Risk factors in Cancer Heredity (SEARCH) and Rotterdam Breast Cancer Study (RBCS), both previously described in Lesueur et al. [29]. Controls were selected from the EPIC-Norfolk cohort study, a population-based cohort study of diet and health based in the same geographical region as SEARCH, together with additional SEARCH controls recruited through general practices in East Anglian region. Genotyping in SEARCH and RBCS was performed by 5′ exonuclease assay (Taqman) using the ABI Prism 7900HT sequence detection system according to the manufacturer's instructions. Primers and probes were supplied directly by Applied Biosystems as Assays-By-Design. Assays included at least two negative controls and 2–5% duplicates per plate.

## Results

Quantile–quantile plots generated from the analyses of individual datasets showed no systematic inflation in the test statistics, indicating no evidence of confounding due to use of non-matched population controls or differential genotyping in cases and controls (Supplementary Figures 2–4, Supplementary File 2). Principal component analysis scatter plots, coloured according to sample source for individual datasets, suggested that the samples were homogeneous within each population (Supplementary File 3). Results obtained using the additive model are presented and discussed below. Corresponding results for the dominant and recessive models are available as online-only supplementary material (Supplementary File 2).

Genomic inflation factors $\lambda$ ($\lambda_{GC}$) for the Swedish, Finnish, CGEMS and the combined datasets after adjusting for were estimated to be 1.013, 1.024, 1.005 and 1.014, respectively. However, there was an excess of significant associations overall at the $10^{-6}$ level, compared with the proportion expected by chance, indicating the presence of susceptibility variants (Fig. 1). These SNPs occurred in three regions known to be associated with breast cancer

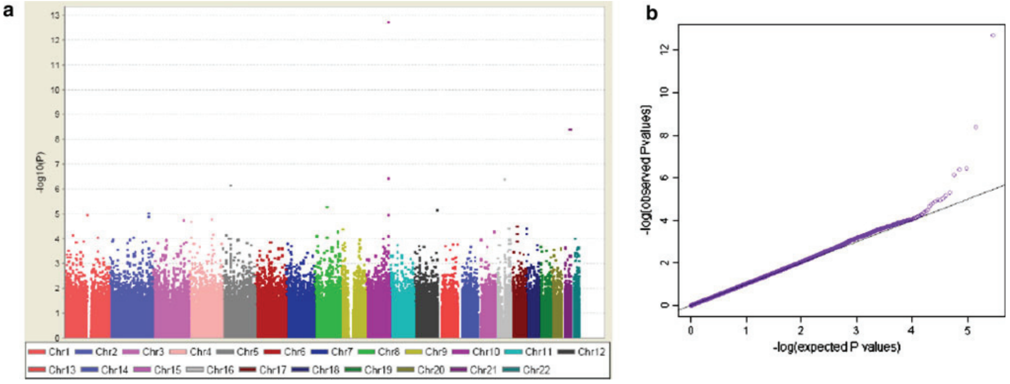


**Fig. 1** Genome-wide association results comparing 2,702 cases and 5,726 controls: **a** Manhattan and **b** quantile–quantile plots of −log10 tranformed *P* values of 285,984 SNPS genotyped

through previous GWAS: 10q26 (FGFR2), 16q12 (TOX3) and 5p12 (MRPS30). P values of single SNP trend tests using the combined GWAS (Swedish, Finnish and CGEMS data) are summarised in a Manhattan plot (Fig. 1). The quantile–quantile plots after exclusion of known susceptibility loci showed no evidence of an excess of associated SNPs over that expected by chance (Fig. 2), consistent with the hypothesis that the known loci include the strongest loci detectable by this platform.

We found strong associations for eight of the known breast cancer susceptibility loci (Table 2; Supplementary File 4). SNPs in 10q26 (FGFR2) [1, 3], 16q12 (TOX3, formerly known as TNRC9) [1], 5p12 (MRPS30) [6], 2q35 [5], 1p11.2 [4], 8q24 [30], 3p (SLC4A7|NEK10) [2] and 14q24.1 (RAD51L1) [4] were found to be strongly associated. The strongest association was for SNP rs1219648 ($P$ trend = 1.93E−13, $\mathrm{OR}_{\mathrm{combined}}$ per allele [95% CI] = 1.32 [1.22–1.42], located within the FGFR2 gene. Originally identified as a susceptibility locus by Easton et al. [1] and by Hunter et al. [3], rs1219648 was also found to be strongly associated within the Swedish ($P$ trend = 1.79E−05, $\mathrm{OR}_{\mathrm{combined}}$ per allele = 1.32 [1.17–1.51]) and Finnish datasets ($P$ trend = 2.00E−04, $\mathrm{OR}_{\mathrm{combined}}$ per allele = 1.30 [1.13–1.49)). The SNP markers located at 1p11.2 (rs11249433), 5p12 (rs7716600, rs4866929 and rs4415084) and RAD51L1 (rs999737) were found to be strongly associated with breast cancer risk in the CGEMS population. The effects of these SNPs were weaker in the Swedish and Finnish populations, although the direction of the effects was consistent. On the other hand, SNPs rs3803662 (TOX3) and rs4973768 (SLC4A7|NEK10) were found to be more significantly associated,with larger effect sizes, within the Swedish and Finnish populations than in the CGEMS population, even though the sample sizes of each were smaller than that of CGEMS. SNP markers located on 2q35 and 8q24 exhibited were found to be significant at the 5% level in the Swedish and CGEMS populations, and achieved $P$ values of 9.23E−06 (OR per allele [95% CI] = 0.85 [0.79–0.91]) and 5.29E−05 (OR per allele = 1.17 [1.08–1.25]), respectively, in the combined analysis. Other regions previously found to be associated with breast cancer, namely, MAP3K1 [1], LSP1 [1], RAD51L1 [4], COX11 [2] and ESR1 [7], did not have appropriate SNPs typed in this study and were thus not examined.

After exclusion of the known loci, six SNPs with the strongest evidence for association, as judged by their one degree of freedom $P$ values for trend, were selected for replication in the SEARCH and RBCS studies (Supplementary File 2). One of the six SNPs, rs4074770 could not be designed, and was not replaced. In the replication stage, none of the SNPs were significant at the 5% level (Table 3). The most significant association was for rs7637164 ($P$ trend = 0.129, $\mathrm{OR}_{\mathrm{combined}}$ per allele = 1.05 [0.99–1.11]); however, this association was in the opposite direction to that observed in the first stage. We conclude that all associations observed in the discovery stage were likely to be false positive associations.

As an alternative approach to detect disease associated SNPs, we performed analyses in which SNPs were classified by the gene footprint into which they lay and genes were classified by pathway. Table 4 summarises pathways for which there was evidence of an overrepresentation of significant SNPs. Full description of association results for SNPs within the significantly associated pathways are presented in Supplementary File 5. Empirical $P$ values for all 212 KEGG pathways are available in Supplementary File 6 as online-only material. Under the broad class of cellular processes, regulation of cytoskeleton (hsa04810, $P =$

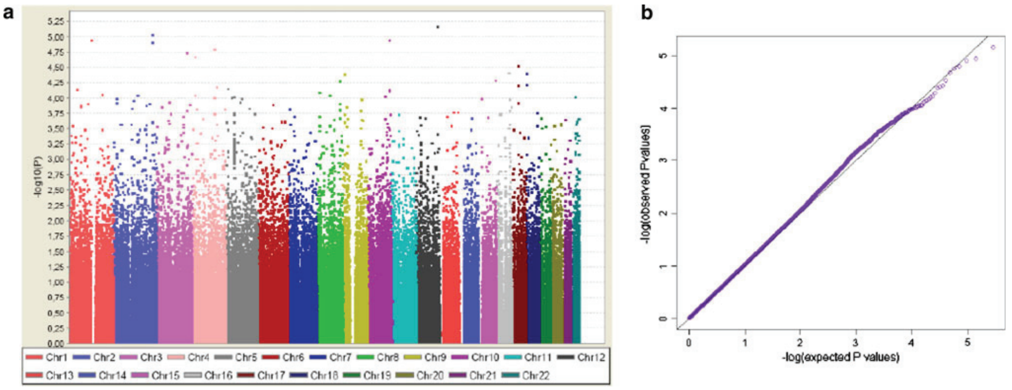


**Fig. 2** Genome-wide association results comparing 2,702 cases and 5,726 controls: **a** Manhattan and **b** quantile–quantile plots of −log10 tranformed $P$ values after removal of previously reported loci

Table 2 Summary of association results of previously reported loci

| Cytoband | Candidate gene | SNP | A1 | MAF | *P* trend, OR per allele (95% CI) | | | | | | Combined | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Swedish | | Finnish | | CGEMS | | Cases | Controls | *P* trend | OR per allele (95% CI) |
| 10q26.13 | FGFR2 | rs1219648 | G | 0.42 | 1.79E−05 | 1.32 (1.17–1.51) | 2.00E−04 | 1.30 (1.13–1.49) | 3.30E−06 | 1.32 (1.18–1.49) | 2702 | 5719 | 1.93E−13 | 1.32 (1.22–1.42) |
| 16q12a1 | TOX3 | rs3803662 | A | 0.30 | 6.14E−04 | 1.27 (1.11–1.46) | 3.21E−04 | 1.29 (1.12–1.48) | 6.53E−02 | 1.13 (0.99–1.28) | 2702 | 5721 | 4.06E−07 | 1.22 (1.13–1.32) |
| 5p12 | MRPS30 | rs7716600 | A | 0.23 | 6.51E−04 | 1.28 (1.11–1.48) | 1.28E−02 | 1.22 (1.04–1.42) | 7.16E−03 | 1.21 (1.05–1.39) | 2697 | 5720 | 7.06E−07 | 1.24 (1.14–1.34) |
| 2q35 | | rs13387042 | G | 0.47 | 7.74E−03 | 0.84 (0.74–0.96) | 5.29E−02 | 0.88 (0.77–1.00) | 2.38E−03 | 0.84 (0.74–0.94) | 2700 | 5721 | 9.23E−06 | 0.85 (0.79–0.91) |
| 1p11.2 | | rs11249433 | G | 0.38 | 1.11E−02 | 1.18 (1.04–1.34) | 1.43E−01 | 1.11 (0.97–1.27) | 5.77E−04 | 1.23 (1.09–1.39) | 2692 | 5695 | 1.13E−05 | 1.18 (1.10–1.27) |
| 8q24 | | rs672888 | G | 0.36 | 2.09E−03 | 1.22 (1.08–1.39) | 3.43E−01 | 1.07 (0.93–1.24) | 4.98E−03 | 1.18 (1.05–1.33) | 2702 | 5723 | 5.29E−05 | 1.17 (1.08–1.25) |
| 5p12 | | rs4866929 | A | 0.49 | 7.51E−02 | 0.89 (0.79–1.01) | 7.74E−01 | 1.02 (0.89–1.17) | 1.26E−05 | 0.77 (0.69–0.87) | 2702 | 5725 | 9.60E−05 | 1.15 (1.07–1.24) |
| 3p | SLC4A7\|NEK10 | rs4973768 | A | 0.46 | 2.81E−02 | 1.15 (1.02–1.31) | 1.72E−03 | 1.24 (1.08–1.42) | 1.60E−01 | 1.09 (0.97–1.22) | 2699 | 5714 | 1.41E−04 | 1.15 (1.07–1.24) |
| 5p12 | | rs4415084 | A | 0.40 | 1.85E−02 | 1.17 (1.03–1.32) | 1.95E−01 | 1.09 (0.96–1.25) | 6.09E−03 | 1.18 (1.05–1.33) | 2699 | 5723 | 1.74E−04 | 1.15 (1.07–1.24) |
| 14q24.1 | RAD51L1 | rs999737 | A | 0.21 | 2.21E−01 | 0.91 (0.78–1.06) | 4.00E−01 | 0.93 (0.78–1.10) | 1.75E−02 | 0.85 (0.74–0.97) | 2701 | 5726 | 8.30E−03 | 0.89 (0.81–0.97) |

Table 3 Summary of association results for the Swedish-only, Finnish-only, CGEMS-only, combined and validation results for selected SNPs using SEARCH and RBCS

| SNP | CHR | A1/A2 | Swedish | | Finnish | | CGEMS | | Combined | | Validation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OR (95% CI) | *P* | OR (95% CI) | *P* | OR (95% CI) | *P* | OR (95% CI) | *P* | Cases/Controls | OR (95% CI) | *P* |
| rs514802 | 8 | A/C | 0.79 (0.70–0.91) | 6.19E−04 | 0.86 (0.74–1.00) | 4.82E−02 | 0.86 (0.77–0.97) | 1.29E−02 | 0.84 (0.78–0.90) | 4.96E−06 | 7,386/7,571 | 1.02 (0.98–1.07) | 0.303 |
| rs6489171 | 12 | G/A | 0.89 (0.77–1.02) | 9.03E−02 | 0.77 (0.67–0.88) | 2.00E−04 | 0.85 (0.75–0.97) | 1.67E−02 | 0.84 (0.77–0.90) | 6.88E−06 | 7,365/7,576 | 0.98 (0.93–1.03) | 0.469 |
| rs724950 | 4 | G/A | 0.84 (0.74–0.95) | 6.59E−03 | 0.89 (0.78–1.03) | 1.08E−01 | 0.84 (0.75–0.94) | 2.79E−03 | 0.85 (0.79–0.92) | 1.59E−05 | 7,360/7,554 | 0.99 (0.95–1.04) | 0.740 |
| rs7637164 | 3 | A/G | 0.79 (0.67–0.93) | 6.00E−03 | 0.88 (0.72–1.08) | 2.28E−01 | 0.77 (0.65–0.90) | 1.38E−03 | 0.80 (0.72–0.89) | 1.80E−05 | 7,277/7,473 | 1.05 (0.99–1.11) | 0.129 |
| rs2697705 | 4 | G/A | 1.31 (1.13–1.51) | 3.16E−04 | 1.16 (0.99–1.36) | 6.83E−02 | 1.15 (1.00–1.32) | 5.15E−02 | 1.20 (1.11–1.31) | 2.12E−05 | 7,305/7,519 | 1.03 (0.97–1.09) | 0.310 |
| rs4074770 | 17 | G/A | 1.19 (1.02–1.38) | 2.29E−02 | 1.12 (0.94–1.34) | 2.10E−01 | 1.29 (1.12–1.48) | 5.17E−04 | 1.21 (1.11–1.32) | 3.02E−05 | – | – | – |

**Table 4** Top ranking pathways of genome-wide pathway analysis results using SNP ratio test

| KEGG pathway ID | Name | Class | No. of SNPs $P < 0.05$ | No. of SNPs in pathway | Number of significantly associated SNPs with $P$ | | | | $P$ | $Q$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | <E−05 | E−04 | E−03 | E−02 | | |
| hsa04810 | Regulation of actin cytoskeleton | Cellular processes; cell motility | 129 | 1870 | 1 | 9 | 26 | 93 | 0.011 | 0.78 |
| hsa00511 | Other glycan degradation | Glycan biosynthesis and metabolism | 11 | 62 | 0 | 0 | 3 | 8 | 0.012 | 0.78 |
| hsa00592 | Alpha-linolenic acid metabolism | Lipid metabolism | 11 | 71 | 0 | 1 | 3 | 7 | 0.018 | 0.78 |
| hsa04710 | Circadian rhythm | Cellular processes; behaviour | 13 | 89 | 0 | 3 | 3 | 7 | 0.020 | 0.78 |
| hsa04640 | Hematopoietic cell lineage | Cellular processes; immune system | 36 | 404 | 0 | 1 | 8 | 27 | 0.021 | 0.78 |
| hsa00983 | Drug metabolism—other enzymes | Xenobiotics biodegradation and metabolism | 25 | 227 | 0 | 1 | 4 | 20 | 0.022 | 0.78 |
| hsa00150 | Androgen and oestrogen metabolism | Lipid metabolism | 19 | 199 | 0 | 0 | 5 | 14 | 0.051 | 0.98 |
| hsa00040 | Pentose and glucuronate interconversions | Carbohydrate metabolism | 9 | 62 | 0 | 0 | 1 | 8 | 0.055 | 0.98 |
| hsa00562 | Inositol phosphate metabolism | Carbohydrate metabolism | 44 | 584 | 0 | 1 | 8 | 35 | 0.065 | 0.98 |
| hsa05216 | Thyroid cancer | Human diseases; cancers | 20 | 217 | 0 | 1 | 5 | 14 | 0.066 | 0.98 |

*KEGG Pathway ID* identifier for pathway curated in the Kyoto Encyclopedia of Genes and Genomes (KEGG), *Name* name of pathway,*Class* class of pathway, *No. of SNPs P < 0.05* number of SNPs with $P < 0.05$ (combined analysis under the additive model), *No. of SNPs in pathway* number of SNPs in KEGG pathway used in the analysis, *No. of significantly associated SNPs with P* number of SNPs with $P <$ specified magnitude, $P$ empirical $P$ value based on comparisons to ratios in datasets where the assignment of case/control status has been randomised to assess the enrichment of significant SNP associations in a pathway context, $Q$ false discovery rate ($Q$ value) computed using the smoother method

0.011), circadian rhythm (hsa04710, $P = 0.020$) and hematopoietic cell lineage (hsa04640, $P = 0.021$) were found to be associated with breast cancer at the significance level of 5%. Pathways related to metabolism of alpha-linolenic acid (hsa00592, $P = 0.018$), glycans (hsa00511, $P = 0.012$), xenobiotics (hsa00983, $P = 0.022$) and steroid hormones (hsa00150, $P = 0.051$) were also found to be significantly associated. $Q$ values for pathways found to be significant ranged from 0.78 to 0.98 for pathways featured in Table 4.

## Discussion

Our combined analysis of three GWAS confirmed evidence for the majority of the known susceptibility alleles (Fig. 1; Table 2; Supplementary File 4). The additive model which has been deemed adequate and recommended for initial GWAS screening [31], is presented in the main article of this paper. That some SNPs were found to be more strongly associated in the CGEMS population, and the effects of others were more prominent in the Swedish and Finnish populations, suggests the fact that each GWAS has limited power to detect variants with small effect sizes (OR per allele between 1.1 and 1.3) [32]. However, the variation in effect sizes by dataset could also be due to variations in sample size, or characteristics specific to each study population. For example, the fact that 210 premenopausal women and 194 of unknown menopausal status were included in the Finnish dataset may help to explain the slight differences in effect sizes observed when comparing the results of this dataset to the Swedish and CGEMS datasets, which included only postmenopausal women. Adjustments for such variables were not made as phenotypic information is not available for all sample populations. The strongest associations found in this combined analysis were all in known loci, notably FGFR2, TOX3 and MRPS30. After removal of loci previously reported to be associated with breast cancer and other variants in LD with these loci, the quantile–quantile plot of 285,973 SNPs interrogated in this study exhibited no departure from the null (Fig. 2), and suggest strongly that loci of similar effect size are unlikely to be found by GWAS. We attempted to validate six SNPs from novel regions with the strongest evidence of association in a larger replication study, but none of these associations could be replicated.

It has been hypothesised that modest associations might be detectable by analyzing pre-defined, biologically meaningful pathways. Given the limited power of our GWAS to detect single SNP associations, we conducted a permutation-based pathway analysis. While there are several alternatives available for the annotation of pathways, of which the most commonly used are Gene Ontology and KEGG Orthology, the latter was chosen to define the pathways examined in this study because of its well-defined levels and terms that correspond to known pathways [33]. Several pathways exhibited an excess of significant associations, over the proportion predicted by chance. However, even though there is still no standard method to deal with the problem of multiple testing in pathway analysis where the dependence structure is unknown, the corresponding local false discovery rates ($Q$ values) were statistically unimpressive (0.78–0.98) for pathways featured in Table 4. It is thus important to emphasise any excess associations may have arisen by chance, and will require validation in additional case–control series.

Gene expression studies have found pathways related to oestrogen signalling [34], circadian rhythm [35] and alpha-linolenic acid metabolism [36] to be associated with breast cancer, corroborating the results of our genome-wide pathway analysis. In addition, Gohlke et al. [37] found key regulatory pathways related to linolenic acid metabolism, drug metabolism, androgen and oestrogen metabolism, hematopoietic cell lineage and regulation of actin cytoskeleton to be associated with breast cancer. While many studies have looked at the effects of differential expression of genes in candidate pathways for breast cancer, less attention has been given to the global effect of SNPs within such pathways. An example of the latter endeavour is the significant association found between the combined effects of SNPs in the oestrogen metabolic pathway and breast cancer risk [38].

In this genome-wide search for pathways associated with breast cancer risk, the most strongly associated pathway relates to the regulation of actin cytoskeleton ($P = 0.011$). This result is understandably driven by the well-known breast cancer associated gene FGFR2. While the significance of this association would not survive multiple testing, a close scrutiny of the statistically significant SNPs within the pathway can potentially uncover novel biology behind the disease. For example, SNP rs2912759 (OR per allele [95% CI] = 0.87 [0.80–0.94], $P = 0.000582$) located within the intron of the FGFR2 gene was documented to be associated with the gene expression of SOX4 in a European population. Human SOX4 has been shown to be expressed in the normal breast and breast cancer cells, and changes in SOX4 gene expression has been suggested to play a role in commitment to the differentiated phenotype in the normal and malignant mammary gland [39]. While at the pathway associations are suggestive rather than confirmed at this stage, it might be worthwhile to examine them in further detail.

Several genes in the drug metabolism and androgen and oestrogen metabolism pathways have already been associated with breast cancer risk, for instance, CYP 19A1, AKR1C4, SULT2A1, SULT2B1, UGT1A6-10 and UGT2B4, among others [38]. Other potential candidate genes include the $\beta$-mannosidase (MANBA) and lactase (LCT) genes, intronic SNPs of which were ranked highly in the glycan degradation pathway, phospholipase A2 (PLA2s) genes in the alpha-linolenic acid metabolism pathway, neuronal PAS domain-containing protein 2 (NPAS2) and period circadian protein homolog 1 (PER1) in the circadian rhythm pathway (Supplementary File 5). While most of the above-mentioned genes and/or their surrounding genes have been suggested to be implicated in cancers on the expression level [40–44], SNPs within these genes have not been examined in relation to breast cancer, and thus merit further examination.

## Conclusion

Our study confirmed several established breast cancer susceptibility loci, but found no new candidates. Nevertheless, we identified pathways related to regulation of actin cytoskeleton, glycan degradation, circadian rhythm, hematopoietic cell lineage, and metabolism of alpha-linoleic acid, drug, and androgen and oestrogen to have suggestive associations with breast cancer risk, which merit further research. The potential of GWAS may be further utilised by complementing traditional single-marker data with biological knowledge of pre-defined pathways.

**Acknowledgments** This work was supported by National Institutes of Health [RO1 CA58427], the Agency for Science, Technology and Research (A*STAR; Singapore), the Nordic Cancer Society, and Märit and Hans Rausing's Initiative against Breast Cancer. J Li is a recipient of an A*STAR Graduate Scholarship (Overseas) award. KH was supported by the Swedish Research Council [523-2006-972]. KC was supported by the Swedish Cancer Society [5128-B07-01PAF]. The Helsinki study was supported by the Helsinki University Central Hospital Research Fund, Academy of Finland, the Finnish Cancer Society and the Sigrid Juselius Foundation. We thank Dr. Kirsimari Aaltonen and RN Hanna Jäntti for their help with the patient data. The SEARCH study was funded by Cancer Research UK. DFE is a Principal Research Fellow of Cancer Research UK. We thank the SEARCH team and Eastern Cancer Registry and Information Centre (ECRIC) for recruitment of the SEARCH cases. The RBCS was supported by the Dutch Cancer Society (Grant DDHK 2004-3124). We acknowledge the clinicians from the Rotterdam Family Cancer Clinic who were involved in collecting the RBCS samples: C. Seynaeve, J. Klijn, J. Collee, and R. Oldenburg. The sponsors took no role in the study design, the collection or analysis of the data, the interpretation of the results, the preparation of the manuscript, or the decision to submit the manuscript for publication.

**Conflict of interest** The authors declare that they have no competing interests.

## References

1. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BA (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447(7148):1087–1093. doi:10.1038/nature05887
2. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, Eccles D, Evans DG, Fletcher O, Johnson N, dos Santos Silva I, Peto J, Stratton MR, Rahman N, Jacobs K, Prentice R, Anderson GL, Rajkovic A, Curb JD, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver WR, Bojesen S, Nordestgaard BG, Flyger H, Dork T, Schurmann P, Hillemanns P, Karstens JH, Bogdanova NV, Antonenkova NN, Zalutsky IV, Bermisheva M, Fedorova S, Khusnutdinova E, Kang D, Yoo KY, Noh DY, Ahn SH, Devilee P, van Asperen CJ, Tollenaar RA, Seynaeve C, Garcia-Closas M, Lissowska J, Brinton L, Peplonska B, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, Hopper JL, Southey MC, Smith L, Spurdle AB, Schmidt MK, Broeks A, van Hien RR, Cornelissen S, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Schmutzler RK, Burwinkel B, Bartram CR, Meindl A, Brauch H, Justenhoven C, Hamann U, Chang-Claude J, Hein R, Wang-Gohrke S, Lindblom A, Margolin S, Mannermaa A, Kosma VM, Kataja V, Olson JE, Wang X, Fredericksen Z, Giles GG, Severi G, Baglietto L, English DR, Hankinson SE, Cox DG, Kraft P, Vatten LJ, Hveem K, Kumle M, Sigurdson A, Doody M, Bhatti P, Alexander BH, Hooning MJ, van den Ouweland AM, Oldenburg RA, Schutte M, Hall P, Czene K, Liu J, Li Y, Cox A, Elliott G, Brock I, Reed MW, Shen CY, Yu JC, Hsu GC, Chen ST, Anton-Culver H, Ziogas A, Andrulis IL, Knight JA, Beesley J, Goode EL, Couch F, Chenevix-Trench G, Hoover RN, Ponder BA, Hunter DJ, Pharoah PD, Dunning AM, Chanock SJ, Easton DF (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41(5):585–590. doi:10.1038/ng.354
3. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ (2007) A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39(7):870–874. doi:10.1038/ng2075
4. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF, Jr., Hoover RN, Chanock SJ, Hunter DJ (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (rad51l1). Nat Genet 41(5):579–584. doi:10.1038/ng.353
5. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, Frigge ML, Geller F, Gudbjartsson D, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Jonsson T, von Holst S, Werelius B, Margolin S, Lindblom A, Mayordomo JI, Haiman CA, Kiemeney LA, Johannsson OT, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39(7):865–869. doi:10.1038/ng2064
6. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Sveinsdottir SG, Alexiusdottir K, Saemundsdottir J, Sigurdsson A, Kostic J, Gudmundsson L, Kristjansson K, Masson G, Fackenthal JD, Adebamowo C, Ogundiran T, Olopade OI, Haiman CA, Lindblom A, Mayordomo JI, Kiemeney LA, Gulcher JR, Rafnar T, Thorsteinsdottir U, Johannsson OT, Kong A, Stefansson K (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40(6):703–706. doi:10.1038/ng.131
7. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, Gu K, Fair AM, Cai Q, Lu W, Shu XO (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41(3):324–328. doi:10.1038/ng.318
8. Ghoussaini M, Pharoah PD (2009) Polygenic susceptibility to breast cancer: current state-of-the-art. Future Oncol 5(5):689–701. doi:10.2217/fon.09.29
9. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M (2009) Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet. doi:10.1038/ejhg.2009.115
10. Wang K, Li M, Bucan M (2007) Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81(6). doi:10.1086/522374
11. Magnusson C, Baron J, Persson I, Wolk A, Bergstrom R, Trichopoulos D, Adami HO (1998) Body size in different periods of life and breast cancer risk in post-menopausal women. Int J Cancer 76(1):29–34. doi:10.1002/(SICI)1097-0215(19980330)76:1<29:AID-IJC6>3.0.CO;2-#
12. Rosenberg LU, Einarsdottir K, Friman EI, Wedren S, Dickman PW, Hall P, Magnusson C (2006) Risk factors for hormone receptor-defined breast cancer in postmenopausal women. Cancer Epidemiol Biomarkers Prev 15(12):2482–2488. doi:10.1158/1055-9965.EPI-06-0489

13. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK (2007) Traf1-c5 as a risk locus for rheumatoid arthritis—a genomewide study. N Engl J Med 357(12):1199–1209. doi:10.1056/NEJMoa073491
14. Bilguvar K, Yasuno K, Niemela M, Ruigrok YM, von Und Zu Fraunberg M, van Duijn CM, van den Berg LH, Mane S, Mason CE, Choi M, Gaal E, Bayri Y, Kolb L, Arlier Z, Ravuri S, Ronkainen A, Tajima A, Laakso A, Hata A, Kasuya H, Koivisto T, Rinne J, Ohman J, Breteler MM, Wijmenga C, State MW, Rinkel GJ, Hernesniemi J, Jaaskelainen JE, Palotie A, Inoue I, Lifton RP, Gunel M (2008) Susceptibility loci for intracranial aneurysm in European and Japanese populations. Nat Genet 40(12):1472–1477. doi:10.1038/ng.240
15. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BW, Janssens AC, Wilson JF, Spector T, Martin NG, Pedersen NL, Kyvik KO, Kaprio J, Hofman A, Freimer NB, Jarvelin MR, Gyllensten U, Campbell H, Rudan I, Johansson A, Marroni F, Hayward C, Vitart V, Jonasson I, Pattaro C, Wright A, Hastie N, Pichler I, Hicks AA, Falchi M, Willemsen G, Hottenga JJ, de Geus EJ, Montgomery GW, Whitfield J, Magnusson P, Saharinen J, Perola M, Silander K, Isaacs A, Sijbrands EJ, Uitterlinden AG, Witteman JC, Oostra BA, Elliott P, Ruokonen A, Sabatti C, Gieger C, Meitinger T, Kronenberg F, Doring A, Wichmann HE, Smit JH, McCarthy MI, van Duijn CM, Peltonen L (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41(1):47–55. doi:10.1038/ng.269
16. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Jarvelin MR, Freimer NB, Peltonen L (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41(1):35–46. doi:10.1038/ng.271
17. Leu M, Humphreys K, Surakka I, Rehnberg E, Muilu J, Rosenstrom P, Almgren P, Jaaskelainen J, Lifton RP, Kyvik KO, Kaprio J, Pedersen NL, Palotie A, Hall P, Gronberg H, Groop L, Peltonen L, Palmgren J, Ripatti S (2010) NordicDB: a nordic pool and portal for genome-wide control data. Eur J Hum Genet. doi:10.1038/ejhg.2010.112
18. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH (2006) A genome-wide association study identifies il23r as an inflammatory bowel disease gene. Science 314(5804):1461–1463. doi:10.1126/science.1135245
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. doi:10.1086/519795
20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909. doi:10.1038/ng1847
21. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A (2009) The snp ratio test: pathway analysis of genome-wide association datasets. Bioinformatics 25(20):2762–2763. doi:10.1093/bioinformatics/btp448
22. Kanehisa M, Goto S (2000) Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
23. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100(16):9440–9445. doi:10.1073/pnas.15305091001530509100
24. Storey JD, Tibshirani R (2003) Statistical methods for identifying differentially expressed genes in DNA microarrays. Methods Mol Biol 224:149–157. doi:10.1385/1-59259-364-X:149
25. R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
26. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics 21(2):263–265. doi:10.1093/bioinformatics/bth457bth457
27. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS (2008) Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 36(Web Server issue):W399–W405. doi:10.1093/nar/gkn296
28. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ (2010) SCAN: SNP and copy number annotation. Bioinformatics 26(2):259–262. doi:10.1093/bioinformatics/btp644
29. Lesueur F, Pharoah PD, Laing S, Ahmed S, Jordan C, Smith PL, Luben R, Wareham NJ, Easton DF, Dunning AM, Ponder BA (2005) Allelic association of the human homologue of the mouse modifier ptprj with breast cancer. Hum Mol Genet 14(16):2349–2356. doi:10.1093/hmg/ddi237
30. Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, Blondal T, Witjes JA, Vermeulen SH, Hulsbergen-van de Kaa CA, Swinkels DW, Ploeg M, Cornel EB, Vergunst H, Thorgeirsson TE, Gudbjartsson D, Gudjonsson SA, Thorleifsson G, Kristinsson KT, Mouy M, Snorradottir S, Placidi D, Campagna M, Arici C, Koppova K, Gurzau E, Rudnai P, Kellen E, Polidoro S, Guarrera S, Sacerdote C, Sanchez M, Saez B, Valdivia G, Ryk C, de Verdier P, Lindblom A, Golka K, Bishop DT, Knowles MA, Nikulasson S, Petursdottir V, Jonsson E, Geirsson G, Kristjansson B, Mayordomo JI, Steineck G, Porru S, Buntinx F, Zeegers MP, Fletcher T, Kumar R, Matullo G, Vineis P, Kiltie AE, Gulcher JR, Thorsteinsdottir U, Kong A, Rafnar T, Stefansson K (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet 40(11):1307–1312. doi:10.1038/ng.229
31. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86(1):6–22. doi:10.1016/j.ajhg.2009.11.017
32. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678. doi:10.1038/nature05911
33. Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the kegg orthology (ko) as a controlled vocabulary. Bioinformatics 21(19):3787–3793. doi:10.1093/bioinformatics/bti430
34. Shen R, Chinnaiyan AM, Ghosh D (2008) Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. BMC Med Genomics 1:28. doi:10.1186/1755-8794-1-28
35. Ma S, Kosorok MR (2010) Detection of gene pathways with predictive power for breast cancer prognosis. BMC Bioinformatics 11:1. doi:10.1186/1471-2105-11-1
36. Chopra P, Shin HJ, Kang J (2008) Global gene map for cancer reveals pathway hotspots. In: Proceedings of the IEEE International conference on Bioinformatics and Biomedicine (BIBM), Philadelphia, USA
37. Gohlke JM, Thomas R, Zhang Y, Rosenstein MC, Davis AP, Murphy C, Becker KG, Mattingly CJ, Portier CJ (2009) Genetic

and environmental pathways to complex diseases. BMC Syst Biol 3:46. doi:10.1186/1752-0509-3-46

38. Low YL, Li Y, Humphreys K, Thalamuthu A, Darabi H, Wedren S, Bonnard C, Czene K, Iles MM, Heikkinen T, Aittomaki K, Blomqvist C, Nevanlinna H, Hall P, Liu ET, Liu J (2010) Multi-variant pathway association analysis reveals the importance of genetic determinants of estrogen metabolism in breast and endometrial cancer susceptibility. PLoS Genet 6:e1001012. doi: 10.1371/journal.pgen.1001012
39. Graham JD, Hunt SM, Tran N, Clarke CL (1999) Regulation of the expression and activity by progestins of a member of the sox gene family of transcriptional modulators. J Mol Endocrinol 22(3):295–304
40. Sud N, Sharma R, Ray R, Chattopadhyay T, Ralhan R (2004) Differential expression of beta mannosidase in human esophageal cancer. Int J Cancer 112(5):905–907. doi:10.1002/ijc.20469
41. Curran JE, Weinstein SR, Griffiths LR (2002) Polymorphic variants of nfkb1 and its inhibitory protein nfkbia, and their involvement in sporadic breast cancer. Cancer Lett 188(1–2): 103–107. doi:S0304383502004603[pii]
42. Yamashita S, Yamashita J, Sakamoto K, Inada K, Nakashima Y, Murata K, Saishoji T, Nomura K, Ogawa M (1993) Increased expression of membrane-associated phospholipase a2 shows malignant potential of human breast cancer cells. Cancer 71(10):3058–3064
43. Zhu Y, Stevens RG, Leaderer D, Hoffman A, Holford T, Zhang Y, Brown HN, Zheng T (2008) Non-synonymous polymorphisms in the circadian gene npas2 and breast cancer risk. Breast Cancer Res Treat 107(3):421–425. doi:10.1007/s10549-007-9565-0
44. Chen ST, Choo KB, Hou MF, Yeh KT, Kuo SJ, Chang JG (2005) Deregulated expression of the per1, per2 and per3 genes in breast cancers. Carcinogenesis 26(7):1241–1246. doi:10.1093/carcin/bgi075

# II

Breast Cancer RESEARCH

RESEARCH ARTICLE Open Access

# A genome-wide association scan on estrogen receptor-negative breast cancer

Jingmei Li[1,2], Keith Humphreys[1], Hatef Darabi[1], Gustaf Rosin[3], Ulf Hannelius[1], Tuomas Heikkinen[4], Kristiina Aittomäki[5], Carl Blomqvist[6], Paul DP Pharoah[7,8], Alison M Dunning[8], Shahana Ahmed[8], Maartje J Hooning[9], Antoinette Hollestelle[10], Rogier A Oldenburg[11], Lars Alfredsson[12], Aarno Palotie[13,14,15,16], Leena Peltonen-Palotie[13,14,15,16], Astrid Irwanto[2], Hui Qi Low[2], Garrett HK Teoh[2], Anbupalam Thalamuthu[2], Juha Kere[3,17,18,19], Mauro D'Amato[3], Douglas F Easton[7,8], Heli Nevanlinna[4], Jianjun Liu[2*], Kamila Czene[1], Per Hall[1*]

## Abstract

**Introduction:** Breast cancer is a heterogeneous disease and may be characterized on the basis of whether estrogen receptors (ER) are expressed in the tumour cells. ER status of breast cancer is important clinically, and is used both as a prognostic indicator and treatment predictor. In this study, we focused on identifying genetic markers associated with ER-negative breast cancer risk.

**Methods:** We conducted a genome-wide association analysis of 285,984 single nucleotide polymorphisms (SNPs) genotyped in 617 ER-negative breast cancer cases and 4,583 controls. We also conducted a genome-wide pathway analysis on the discovery dataset using permutation-based tests on pre-defined pathways. The extent of shared polygenic variation between ER-negative and ER-positive breast cancers was assessed by relating risk scores, derived using ER-positive breast cancer samples, to disease state in independent, ER-negative breast cancer cases.

**Results:** Association with ER-negative breast cancer was not validated for any of the five most strongly associated SNPs followed up in independent studies (1,011 ER-negative breast cancer cases, 7,604 controls). However, an excess of small *P*-values for SNPs with known regulatory functions in cancer-related pathways was found (global *P* = 0.052). We found no evidence to suggest that ER-negative breast cancer shares a polygenic basis to disease with ER-positive breast cancer.

**Conclusions:** ER-negative breast cancer is a distinct breast cancer subtype that merits independent analyses. Given the clinical importance of this phenotype and the likelihood that genetic effect sizes are small, greater sample sizes and further studies are required to understand the etiology of ER-negative breast cancers.

## Introduction

Breast cancer is a heterogeneous disease and can be characterized on the basis of estrogen receptor (ER) expression in the tumour cells. The two breast cancer subtypes (ER-positive and ER-negative) are generally considered as biologically distinct diseases and have been associated with remarkably different gene expression profiles [1,2]. ER status is important clinically, and is used both as a prognostic indicator and treatment predictor since it determines if a patient may benefit from anti-estrogen therapy. Approximately one-third of all breast cancers are ER-negative, and cancers of this ER subtype are highly age-dependent and generally have a more aggressive clinical course than hormone receptor-positive disease.

Estimates show that close to a third of the total risk of breast cancer may be attributed to heritable factors [3]. Several large-scale genome-wide single nucleotide polymorphism (SNP) association studies (GWAS) have identified multiple susceptibility loci for breast cancer [4-11], but it is estimated that the currently known common risk variants identified by this approach explains only 5.8% of the proportion of familial risk of breast cancer.

* Correspondence: liuj3@gis.a-star.edu.sg; per.hall@ki.se
[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, P.O. Box 281, Stockholm 17177, Sweden
[2]Human Genetics, Genome Institute of Singapore, 60 Biopolis St, Singapore 138672, Singapore
Full list of author information is available at the end of the article

Aside from traditional agnostic SNP studies, pathway-based approaches have also emerged in the recent GWAS literature [12-20]. These novel methods have been developed to mine modest association signals from genome-wide SNP data using prior knowledge on biologically pathways and networks, and have the potential to complement traditional agnostic SNP approaches to provide fertile grounds for follow-up studies of both a genetic and molecular nature. Subtypes of breast cancer, to our knowledge, have not been studied using a pathway-based approach.

Although many of the SNPs identified for breast cancer through GWAS scans have been found to be more strongly associated with ER-positive disease than ER-negative disease [21,22], there is no quantitative assessment on whether breast cancers of the two different ER subtypes share a polygenic component. In this study, we performed a genome-wide association scan on 617 ER-negative cases and 4,583 controls, the first of its kind, and examined 285,984 SNPs for common variants and biological pathways associated with this unique subtype of breast cancer. We also searched for evidence that ER-negative breast cancer is distinct from ER-positive breast cancer by assessing the amount of shared polygenic variation between the two breast cancer subtypes.

## Materials and methods

Full methods accompany this paper in Additional file 1.

### Study populations used in the discovery stage

Table 1 summarizes the demographics of cases and controls used in this study. The discovery stage consists of cases and controls from Finland and Sweden. The validation stage consists of breast cancer cases from two further studies: the Study of Epidemiology and Risk factors in Cancer Heredity (SEARCH) and Rotterdam Breast Cancer Study (RBCS) (1,011 ER-negative cases, 7,604 controls), both previously described in Lesueur *et al.* [23]. Informed consent was obtained from all subjects. For all populations, blood samples were obtained from individuals according to protocols and informed-consent procedures approved by institutional review boards.

Briefly, the Swedish sample set included subjects who were drawn from a parent population-based case control study of postmenopausal breast cancer which has been described elsewhere [24,25]. Case subjects were women born in Sweden who were 50 to 74 years of age at diagnosis and diagnosed with breast cancer between October 1993 and March 1995. A total of 803 individuals diagnosed with invasive breast cancer and with available blood samples were selected for GWAS genotyping in an independent GWAS looking at overall breast cancer risk [26]. Of these women, 153 individuals were diagnosed with the ER-negative disease and were included in the present study. In addition, a total of 1,414 Swedish controls were included from the parent study and an additional Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study [27].

The Finnish breast cancer study population consists of two series of unselected breast cancer patients and additional familial cases ascertained at the Helsinki University Central Hospital. The first series of patients was collected in 1997 to 1998 and 2000 and covers 79% of all consecutive, newly diagnosed cases during the collection periods [28,29]. The second series, containing newly diagnosed patients, was collected in 2001 to 2004 and covers 87% of all such patients treated at the hospital during the collection period [30]. The collection of additional familial cases has been described previously [31]. We genotyped a total of 782 breast cancer cases in an independent GWAS for overall breast cancer risk

**Table 1 Summary of samples and genotyping platforms used in the discovery and validation stages**

| Stage | Study | Type | No. of samples after quality control | Genotyping platform |
|---|---|---|---|---|
| Discovery | Swedish | ER-negative cases | 153 | HumanHap300 supplemented by HumanHap240S |
| | | Controls | 764 | HumanHap550 |
| | | Additional controls from EIRA study | 650 | HumanHap300 |
| | Finnish | ER-negative cases | 226 | HumanHap550 |
| | | ER-negative cases | 238 | Quad610 (v1) |
| | | Controls | 3169 | HumanHap370Duo |
| Validation | SEARCH and RBCS | ER-negative cases | 1011 | Taqman |
| | | Controls | 7604 | Taqman |

ER, estrogen receptor; RBCS, Rotterdam Breast Cancer Study; SEARCH, Study of Epidemiology and Risk factors in Cancer Heredity.

[26], of which 226 ER-negative cases were used in the present study. An additional 238 Finnish ER-negative cases were also genotyped for this study, using a different platform. Of these 464 women with ER-negative breast cancer, 207 were sporadic and 257 were familial breast cancer cases. Population control data were obtained from the Finnish Genome Centre on 3,169 healthy population controls described in [32-35].

SEARCH is a population-based case-control study comprising 7,093 cases identified through the East Anglian Cancer Registry: prevalent cases diagnosed age <55 from 1991 to 1996 and alive when the study started in 1996, and incident cases diagnosed <70 diagnosed after 1996. Controls ($N$ = 8,096) were selected from the EPIC-Norfolk cohort study, a population-based cohort study of diet and health based in the same geographical region as SEARCH, together with additional SEARCH controls recruited through general practices in East Anglian region.

RBCS is a hospital-based case-control study comprising 799 cases characterized as familial breast cancer patients selected from the Rotterdam Family Cancer Clinic at the Erasmus Medical Center, of which 141 are ER-negative. Controls ($N$ = 801) were spouses or mutation-negative siblings of heterozygous Cystic Fibrosis mutation carriers selected from the Department of Clinical Genetics at the Erasmus Medical Center. Both cases and controls were recruited between 1994 and 2006.

### Genotyping and quality control filters

Genotyping for all samples was performed according to the Illumina Infinium 2 assay manual (Illumina, San Diego, CA, USA), as described previously [36]. The genotyping platforms used for this study are listed in Table 1. Apart from the 3,170 Finnish controls which were genotyped on the HumanHap370Duo assay as described previously [32,34], genotyping for all other Finnish and Swedish samples was performed at the Genome Institute of Singapore.

Each dataset was filtered to remove individuals with >10% missing genotypes, and SNPs with >10% missing data, or minor allele frequency (MAF) <0.03, or not in Hardy-Weinberg equilibrium (HWE) ($P$ < 0.05/number of SNPs after quality control) and individual samples with evidence of possible DNA contamination, common ancestry or cryptic family relationships. Quality control was carried out using the software Plink [37]. To account for population outliers and correct for differential ancestry between cases and controls that may exist in the dataset after familial outlier removal, a principal component (PC) analysis was conducted using the EIGENSTRAT software (Broad Institute, Boston, MA, USA) [38].

A total of 617 ER-negative cases and 4,583 controls passed the quality control for samples. The 285,984 SNPs that passed quality control filters in all sample sets were merged into a single file for analysis.

The five most strongly associated SNPs in the combined analysis, which had effects in the same direction for both studies in the discovery stage (Swedish and Finnish) were forwarded for validation in SEARCH and RBCS. Genotyping in SEARCH and RBCS was performed by 5'exonuclease assay (Taqman) using the ABI Prism 7900HT sequence detection system (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's instructions.

All SNP chromosomal positions were based on NCBI Build 36.

### Statistical analysis

Figure 1 gives a broad overview of the analytical strategy for the single marker association analysis and pathway analysis.

#### *Single marker association analysis*

Logistic regression models with genotype coded 0, 1, 2 and treated as a continuous covariate (one at a time), were fitted for each SNP that passed quality control. An additive genetic effect on the logit scale was assumed to characterize the associations. Separate analyses were performed for the Swedish and Finnish datasets as well as a combined analysis.

In the combined analysis, the final model included as covariates the SNP genotype, an indicator variable specifying country (Sweden and Finland), and interaction effects of Eigen values of PCs × country specified in such a way that country-specific PCs were implemented for the relevant subjects. Quantile-quantile plots were used to check for systematic genotyping error or bias due to unaccounted underlying population substructure. Manhattan plots were generated to summarize the -log transformed $P$-values of all SNPs examined.

#### *Pathway analysis using discovery set (Swedish and Finnish samples)*

Pathway analysis of the discovery GWAS dataset was conducted using the SNP ratio test (SRT) SRT was used to investigate the associations with breast cancer for 212 pathways and their genes (approximately 4,700) taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (05/12/08) [39].

To evaluate the association between regulatory SNPs-defined pathways and ER-negative breast cancer, we used the downloadable database from mRNA by SNP Browser [40] to map SNPs, which are significantly associated with gene expression on a genome-wide level (LOD >6), to genes. In total, 7,698 SNPs were mapped to 3,740 probes with a LOD score >6. These 3,740

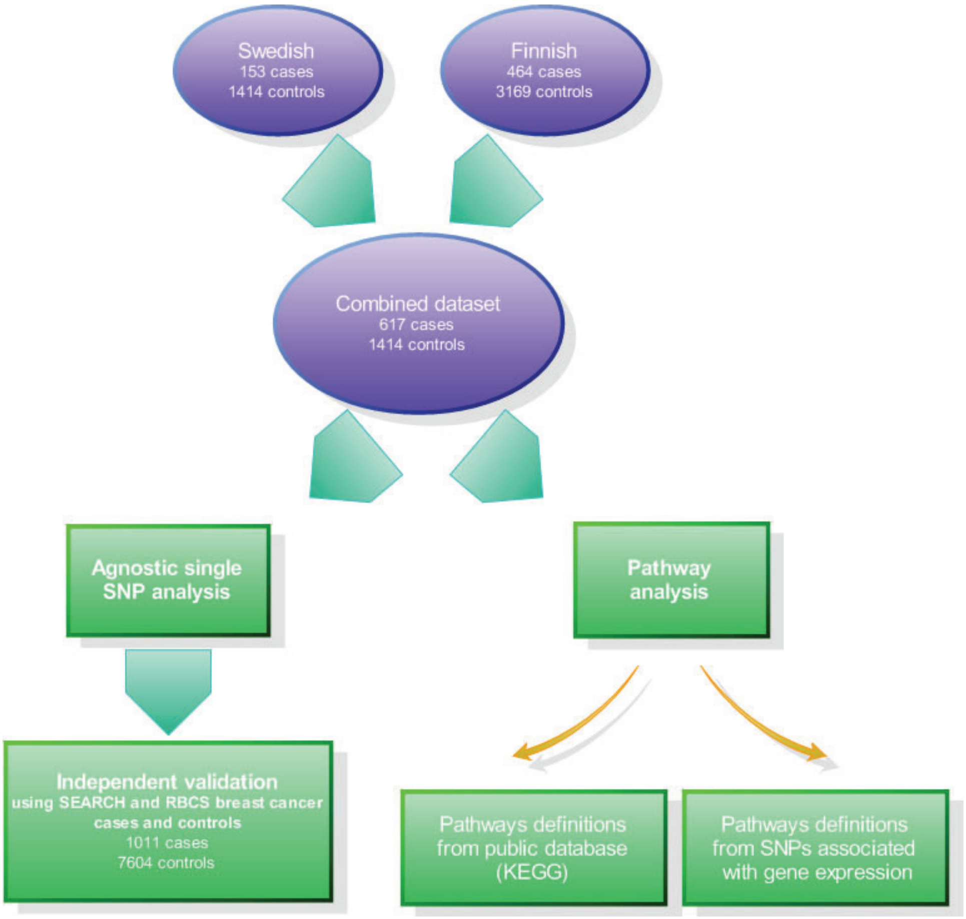

Swedish
153 cases
1414 controls
Finnish
464 cases
3169 controls
Combined dataset
617 cases
1414 controls
Agnostic single SNP analysis
Pathway analysis
Independent validation
using SEARCH and RBCS breast cancer cases and controls
1011 cases
7604 controls
Pathways definitions from public database (KEGG)
Pathways definitions from SNPs associated with gene expression

**Figure 1** Schematic diagram of analytical strategies for agnostic single marker association analysis and pathway analysis.

probes could be mapped to 2,070 genes, and out of these, 554 genes, regulated by 1,720 SNPs, were annotated as belonging to one or several of the 182 KEGG pathways.

Among five regulatory SNP-defined pathways found to be significantly associated with ER-negative breast cancer, four belonged to the pathway class "cancer". To evaluate if the abundance of small *P*-values from regulatory SNPs involved in cancer-related pathways was statistically significant as a whole, we also assessed the departure of the distribution of the trend test statistics from the null distribution, assuming that none of the SNPs was associated with ER-negative breast cancer as an outcome. For this purpose, we performed the "admixture maximum likelihood" test described by Tyrer *et al.* [41] to obtain a global *P*-value for 165 unique SNPs from 15 cancer-related pathways (hsa052*) curated in the KEGG database.

### *Analysis of shared polygenic variation between ER-negative and ER-positive breast cancer subtypes*

We assessed the polygenic component of breast cancer risk using a procedure for creating sample scores which has been described elsewhere [42]. Briefly, ER-positive breast cancer cases and healthy controls from either the Finnish or Swedish study were used as a "training set" to derive a list of SNPs used for scoring in two "target sets", consisting of either ER-positive breast cancer cases and healthy controls or ER-negative breast cancer cases and healthy controls in the other population. Figure 2 gives a broad overview of the analytical strategy for assessing common polygenic variation.

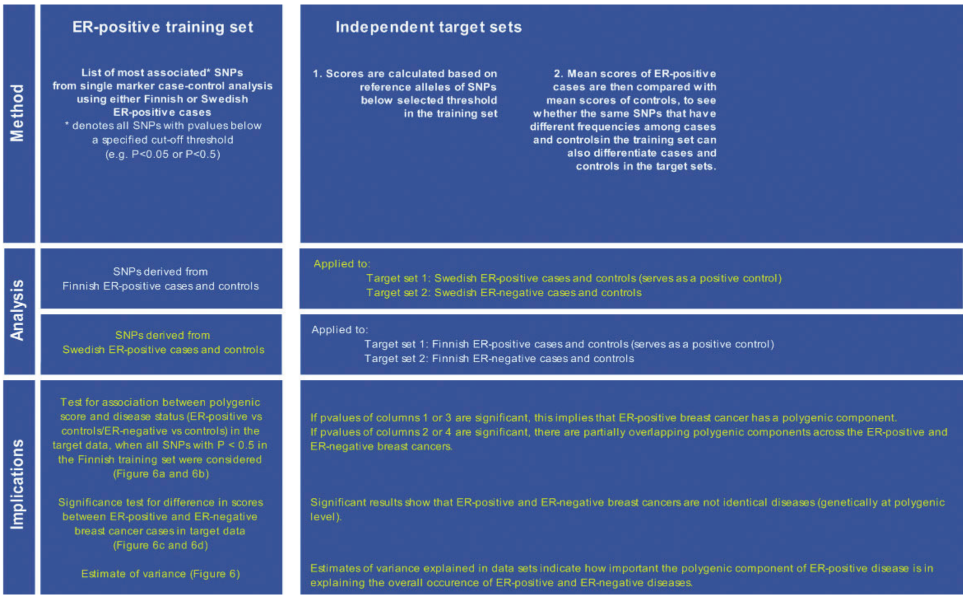
The polygenic score for each individual was calculated by summing the number of score alleles weighed by the log of their odds ratio from the training sample, across all SNPs included in the score. SNPs were included in the score if they achieved a *P*-value less than a particular threshold in the training sample. The "—score" function in Plink [37] was used to calculate scores. To capture association signals with very small effects in the calculation of the polygenic component of the disease, we used non-stringent significance thresholds ($P < 0.01$, $P < 0.05$, $P < 0.10$, $P < 0.20$, $P < 0.30$, $P < 0.40$ and $P < 0.50$). Scores were calculated for the seven *P*-value thresholds.

The extent of shared polygenic variation between ER-positive breast cancers in the training sample and ER-positive and ER-negative breast cancers in the corresponding target samples was assessed by fitting logistic regression models to disease state, as a function of score, in the target samples. Regression models, adjusted for the number of non-missing genotypes, were fitted to assess the differences in the extent of shared polygenic variation (scores) between the ER-positive and ER-negative target samples in case-only analyses.

PLINK (v1.06) [37], SNP Ratio Test [19], R (v2.8.0) [43], Quanto [44], AML [41], Qlikview (v8.5) [45], HaploView [46] and LocusZoom [47] were used for data management, quality control, statistical analyses, and graphics. All reported tests are two-sided.

## Results

In this study, we tested the association of 285,984 loci with ER-negative breast cancer in two independent populations consisting of a total of 617 cases and 4,583 controls. It appears that the overall population substructure was adequately accounted for, since a systematic deviation from the expected distribution was not observed in the quantile-quantile plot (Supplementary Figures 2, 3 and 4 in Additional file 2). Quantile-quantile plots generated from the analyses of individual datasets showed that there was no within-study systematic error arising from the use of non-matched population controls or genotyping at different facilities (Supplementary Figures 2 and 3 in Additional file 2). Genotype cluster plots were examined for SNPs with $P < 10^{-5}$. Manual reclustering was performed for six SNPs with poor genotype cluster plots. SNPs rs4660646 and rs2462692 were



|  | ER-positive training set | Independent target sets |
|---|---|---|
| Method | List of most associated* SNPs from single marker case-control analysis using either Finnish or Swedish ER-positive cases<br>* denotes all SNPs with pvalues below a specified cut-off threshold (e.g. P<0.05 or P<0.5) | 1. Scores are calculated based on reference alleles of SNPs below selected threshold in the training set<br>2. Mean scores of ER-positive cases are then compared with mean scores of controls, to see whether the same SNPs that have different frequencies among cases and controlsin the training set can also differentiate cases and controls in the target sets. |
| Analysis | SNPs derived from Finnish ER-positive cases and controls | Applied to:<br>Target set 1: Swedish ER-positive cases and controls (serves as a positive control)<br>Target set 2: Swedish ER-negative cases and controls |
|  | SNPs derived from Swedish ER-positive cases and controls | Applied to:<br>Target set 1: Finnish ER-positive cases and controls (serves as a positive control)<br>Target set 2: Finnish ER-negative cases and controls |
| Implications | Test for association between polygenic score and disease status (ER-positive vs controls/ER-negative vs controls) in the target data, when all SNPs with P < 0.5 in the Finnish training set were considered (Figure 6a and 6b) | If pvalues of columns 1 or 3 are significant, this implies that ER-positive breast cancer has a polygenic component.<br>If pvalues of columns 2 or 4 are significant, there are partially overlapping polygenic components across the ER-positive and ER-negative breast cancers. |
|  | Significance test for difference in scores between ER-positive and ER-negative breast cancer cases in target data (Figure 6c and 6d) | Significant results show that ER-positive and ER-negative breast cancers are not identical diseases (genetically at polygenic level). |
|  | Estimate of variance (Figure 6) | Estimates of variance explained in data sets indicate how important the polygenic component of ER-positive disease is in explaining the overall occurence of ER-positive and ER-negative diseases. |

**Figure 2** Summary of scoring procedure for assessment of common polygenic variation.

omitted from further analysis as they could not be reclustered. SNPs rs4549482, rs1984492, rs1389545 and rs3748648 were not found to be strongly associated with ER-negative breast cancer after reclustering (Table S1 in Additional file 3).

Figure 3 shows a Manhattan plot summarizing the -log-transformed *P*-values of 285,984 SNPs analyzed in this study. In a combined analysis of individuals of Swedish and Finnish backgrounds, the strongest association with ER-negative breast cancer below the threshold for genome-wide significance was for a locus marked by rs361147 on chromosome 4 (*P* trend = 3.13 × $10^{-13}$; OR $_{\text{per allele}}$ = 0.60) (Table S2 in Additional file 3). This was the only SNP to achieve statistical significance at the genome-wide level (α = 5 × $10^{-8}$). Overall, no significant signal peak was identified in this study (Figures 4, 5, 6, 7, 8).

Nevertheless, we selected five SNPs to be validated in a combined dataset of two independent studies (Table S2 in Additional file 3). SNPs rs7039994 and rs12000794, located 106310 base pairs away from each other on chromosome 9, were found to be in high LD (r2 = 0.797; D' = 0.952). The former was kept and validated in the SEARCH dataset as its associated *P*-value was smaller and it was in closer proximity to coding regions (downstream of *INVS*|*TEX10*). SNP rs3777218 was selected over rs11882068 due to a better regional signal peak. Other SNPs selected for validation included rs361147 as mentioned above, rs6993922, rs4726078 (within transcript of *PRKAG2*), and rs3777218 (within transcript of *RHOBTB3*). Of the five SNPs forwarded for validation, rs4726078 could not be designed and was replaced by rs10952315 (r2 = 0.977 in Centre d'Etude du Polymorphisme Humain (CEPH) from Utah (CEU) HapMap samples). None of the SNPs was significantly associated at the 5% level in the second stage. The smallest *P*-value obtained was for the surrogate rs10952315 (OR 1.02; 95% CI: 0.93 to 1.13).

To analyze our GWAS data in a pathway context we conducted a permutation-based analysis using the KEGG database. Pathways defined by SNPs located within transcript of genes that were found to be significantly associated with ER-negative breast cancer after 1,000 phenotype permutations at a threshold of $P_{\alpha = 0.05}$ < 0.05 (uncorrected) were: pentose and glucuronate interconversions (hsa00040) (*P* = 0.022), starch and sucrose metabolism (hsa00500) (*P* = 0.042), and gap junction (hsa04540) (*P* = 0.037) (Table 2).

In addition, we limited the analysis to pathway definitions involving only known regulatory SNPs [48]. The GWAS SNPs were first mapped to genes, and then subsequently to KEGG pathways based on publicly available

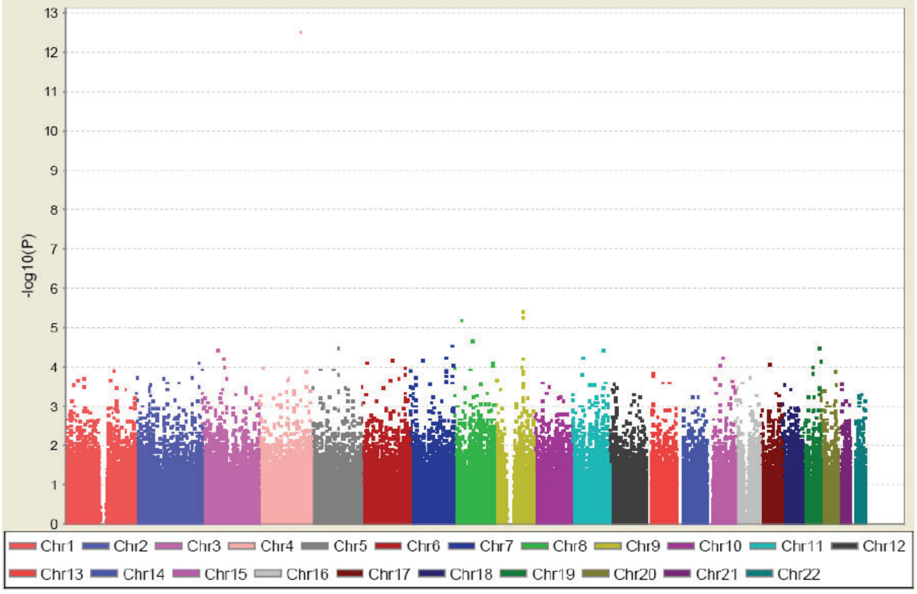


**Figure 3 Genome-wide *P*-values (-log$_{10}$P) of the logistic regression analysis plotted against chromosomal position.**

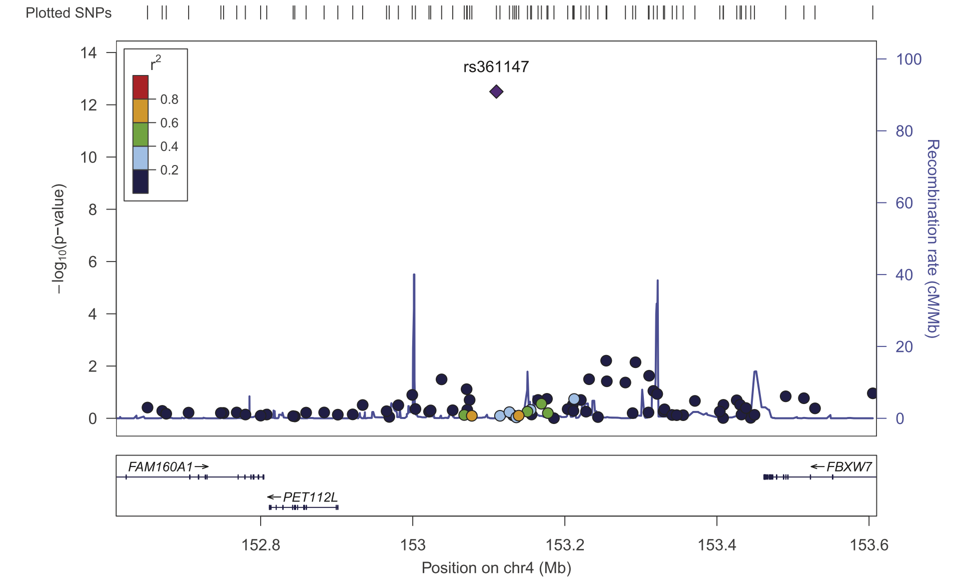


**Figure 4 Plot of regional association signals for rs361147 forwarded for validation**.

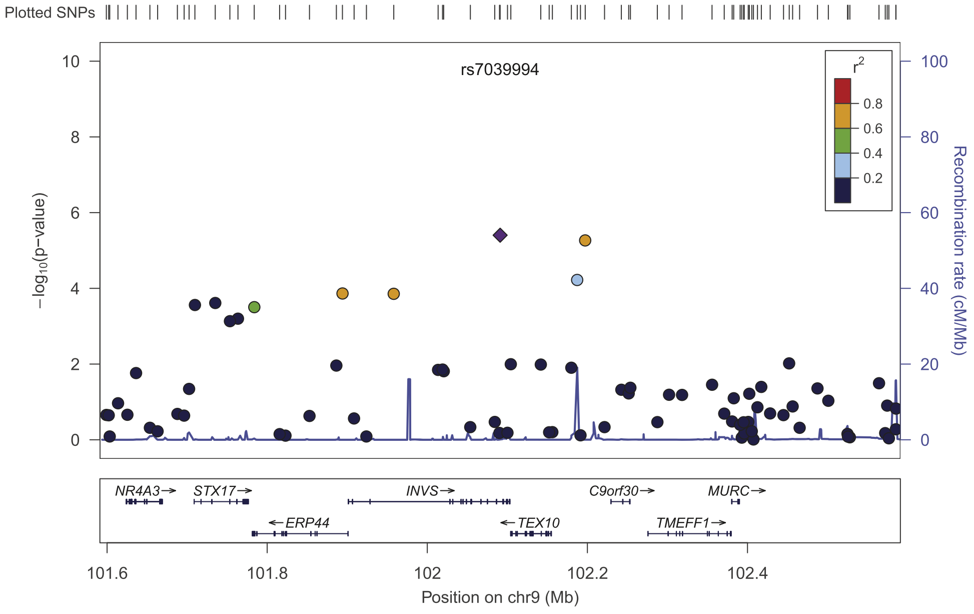


**Figure 5 Plot of regional association signals for rs7039994 forwarded for validation**.

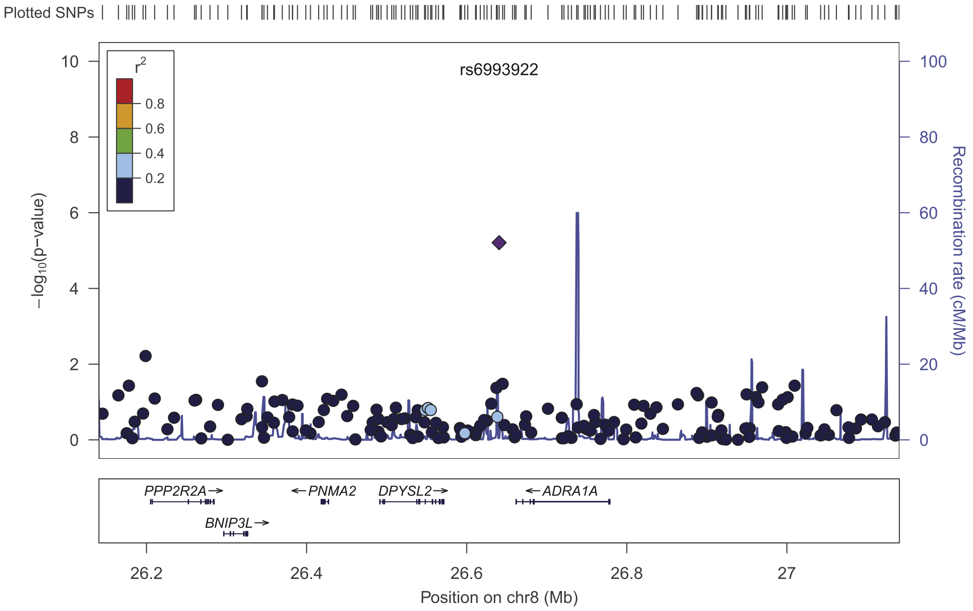


**Figure 6 Plot of regional association signals for rs6993922 forwarded for validation**.

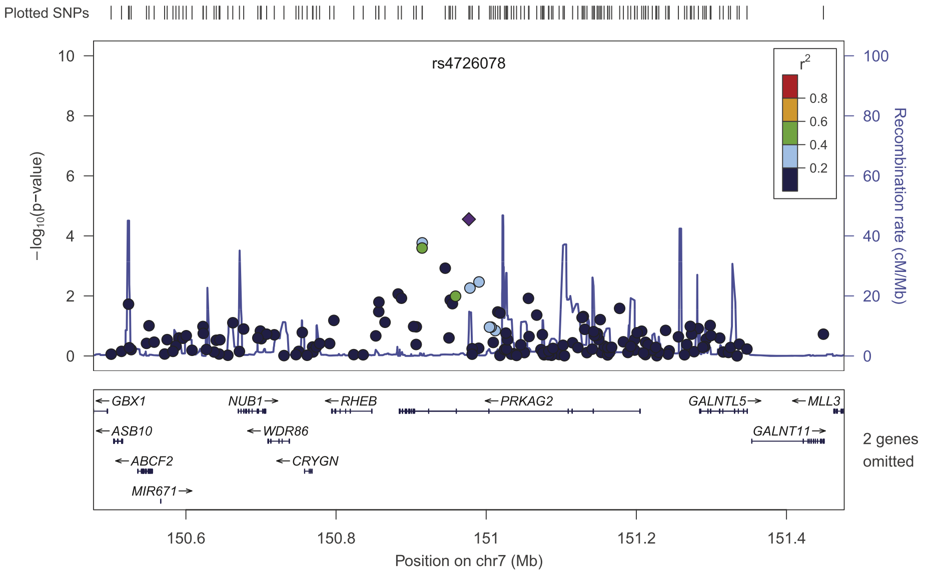


**Figure 7 Plot of regional association signals for rs4726078 forwarded for validation**.

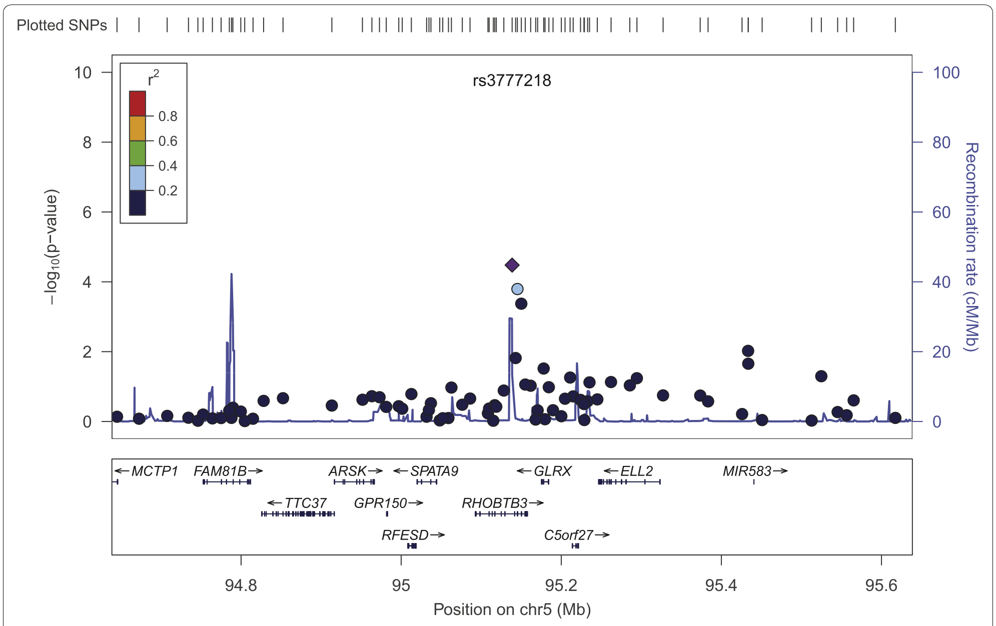


**Figure 8 Plot of regional association signals for rs3777218 forwarded for validation**.

gene regulatory data from lymphoblastoid cells [48]. Only genes with regulatory functions significant on a genome-wide significant level were selected, resulting in 1,720 SNPs regulating members of 182 KEGG pathways being used in our analysis. Pathways that were found to be significant by SRT after 1,000 phenotype permutations at a threshold of $P_{\alpha\ =\ 0.05} < 0.05$ were: long-term potentiation (hsa04720), glioma (hsa05214), non-small cell lung cancer (hsa05223), pancreatic cancer (hsa05212), and prostate cancer (hsa5215) (Table 3). The focal adhesion pathway (hsa04510) was found to be marginally significant ($P_{\alpha\ =\ 0.05}$ = 0.052). Two pathways each tagged by only a single SNP, glyoxylate and dicarboxylate metabolism (hsa00630) and glycosphingolipid biosynthesis - ganglio series (hsa00604), were removed from the evaluation of the final results.

Regulatory SNPs involved in pathways associated with cancer (hsa052*) appeared to be overrepresented by small *P*-values (Figure 9). To evaluate if the combined effect of these signals was statistically significant as a whole, we next carried out a global test of significance for all unique SNPs in the cancer pathways. The AML analysis performed using an algorithm developed by Tyrer *et al.* [41], yielded *P*-values (α = 0.05) of 0.0028 (crude) and 0.052 (adjusted for population stratification).

Figure 10 shows the results of analyses aimed at assessing the shared polygenic component between ER-positive and ER-negative breast cancer. Estimates of variance explained in datasets indicate how important the polygenic component of ER-positive disease is in explaining the overall occurrence of ER-positive and ER-negative diseases. The proportion of variance explained for all categories of *P*-value cut-offs, with the exception of $P < 0.05$ in the Swedish ER-positive target sample, was higher in the ER-positive target datasets than the ER-negative target datasets.

We test for association between polygenic score and disease status (ER-positive vs controls/ER-negative vs controls) in the target data, when seven groups of SNPs with different *P*-values thresholds in the training sets were considered (Figure 10a, b). Due possibly to limited statistical power (Table S3 in Additional file 3), even at the least stringent *P*-value threshold ($P < 0.50$), the ER-positive and ER-negative breast cancer target case-control datasets failed to provide statistically significant evidence of a polygenic component for ER-positive cancer, or evidence of a polygenic component shared between the two cancers, when training was based on the ER-positive training case-control datasets (Figure 10a, b). Nevertheless, when we relaxed the *P*-value cut-off in the

**Table 2 Top ranking pathways of genome-wide pathway analysis results using SNP ratio test (*P* < 0.1)**

| KEGG ID | Pathway name<br>*Class* | No. of SNPs<br>*P* < 0.05 | No. of SNPs in pathway | Number of significantly associated SNPs with *P* | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | E-05 | E-04 | E-03 | E-02 | *P* |
| 00040 | Pentose and glucuronate interconversions<br>*Metabolism; Carbohydrate Metabolism* | 11 | 63 | 0 | 1 | 2 | 8 | 0.022 |
| 04540 | Gap junction<br>*Cellular Processes; Cell Communication* | 95 | 1,366 | 1 | 0 | 16 | 78 | 0.037 |
| 00500 | Starch and sucrose metabolism<br>*Metabolism; Carbohydrate Metabolism* | 22 | 237 | 0 | 2 | 4 | 16 | 0.042 |
| 00604 | Glycosphingolipid biosynthesis ganglio series<br>*Metabolism; Glycan Biosynthesis and Metabolism* | 20 | 216 | 0 | 0 | 4 | 16 | 0.051 |
| 00230 | Purine metabolism<br>*Metabolism; Nucleotide Metabolism* | 106 | 1,618 | 1 | 2 | 16 | 87 | 0.054 |
| 04130 | SNARE interactions in vesicular transport<br>*Genetic Information Processing; Folding, Sorting and Degradation* | 19 | 206 | 0 | 4 | 1 | 14 | 0.060 |
| 03022 | Basal transcription factors<br>*Genetic Information Processing; Transcription* | 11 | 105 | 0 | 0 | 4 | 7 | 0.062 |
| 04910 | Insulin signaling pathway<br>*Cellular Processes; Endocrine System* | 61 | 889 | 2 | 6 | 9 | 44 | 0.071 |
| 04350 | TGF-beta signaling pathway<br>*Environmental Information Processing; Signal Transduction* | 43 | 586 | 0 | 1 | 9 | 33 | 0.077 |
| 04330 | Notch signaling pathway<br>*Environmental Information Processing; Signal Transduction* | 25 | 321 | 0 | 0 | 4 | 21 | 0.087 |
| 04614 | Renin-angiotensin system<br>*Cellular Processes; Endocrine System* | 8 | 78 | 0 | 0 | 1 | 7 | 0.092 |

KEGG ID, Kyoto Encyclopedia of Genes and Genomes pathway identifier (hsa*); P, P-value of permutation test; SNP, single nucleotide polymorphism

training dataset to 0.5, the Swedish ER-positive breast cancer target dataset showed borderline significance for a shared polygenic component with ER-positive breast cancer, based on the Finnish ER-positive training dataset (Figure 10a, *P* = 0.066).

In a separate case-only analysis, we performed a significance test for difference in scores between ER-positive and ER-negative breast cancer cases in the target data. Significant results show that ER-positive and ER-negative breast cancers are not identical diseases (genetically at polygenic level) (Figures 10c, d). The difference in scores between ER-positive and ER-negative samples was found to be statistically significant for all categories of *P*-value cut-offs in the Finnish target case-only samples, with the exception of the most associated SNPs (Figure 10d).

## Discussion

Little is known about the genetic predisposition to estrogen receptor-negative breast cancer. This subtype is characterized by lower age of onset, a more aggressive disease and low or no response to selective estrogen receptor modulators or aromatase inhibitors. We have examined our GWAS data on two different levels: single marker and pathway. We also provided evidence that breast cancer is a heterogeneous disease with a polygenic nature, with significant differences between the polygenic component between ER-positive and ER-

**Table 3 Top ranking pathways of genome-wide pathway analysis using regulatory SNPs**

| Pathway name (KEGG ID)<br>*Class* | SRT P | *P*-value distribution of SNPs<br>*P* < 0.01 | <br>0.01 ≤ *P* < 0.05 | <br>0.05 ≤ *P* < 0.1 | N | *P* of most significant SNP in pathway |
|---|---|---|---|---|---|---|
| Glioma (hsa05214)<br>*Cancers* | 0.0394 | 1 | 5 | 4 | 26 | 0.0028 |
| Long-term potentiation (hsa04720)<br>*Nervous System* | 0.0394 | 0 | 3 | 2 | 16 | 0.0314 |
| Non-small cell lung cancer (hsa05223)<br>*Cancers* | 0.0394 | 1 | 5 | 3 | 24 | 0.0028 |
| Pancreatic cancer (hsa05212)<br>*Cancers* | 0.0413 | 2 | 5 | 3 | 33 | 0.0028 |
| Prostate cancer (hsa05215)<br>*Cancers* | 0.0488 | 3 | 3 | 6 | 32 | 0.0003 |
| Focal adhesion (hsa04510)<br>*Cell Communication* | 0.0525 | 1 | 7 | 9 | 71 | 0.0028 |
| Chemokine signaling pathway (hsa04062)<br>*Immune System* | 0.0582 | 1 | 8 | 7 | 72 | 0.0080 |
| Pathways in cancer (hsa05200)<br>*Cancers* | 0.0582 | 2 | 12 | 15 | 151 | 0.0028 |
| Melanogenesis (hsa04916)<br>*Endocrine System* | 0.0657 | 2 | 2 | 2 | 26 | 0.0003 |
| B cell receptor signaling pathway (hsa04662)<br>*Immune System* | 0.0713 | 0 | 5 | 3 | 29 | 0.0314 |
| GnRH signaling pathway (hsa04912)<br>*Endocrine System* | 0.0732 | 0 | 6 | 6 | 46 | 0.0115 |
| Fc epsilon RI signaling pathway (hsa04664)<br>*Immune System* | 0.0769 | 0 | 6 | 6 | 33 | 0.0314 |
| VEGF signaling pathway (hsa04370)<br>*Signal Transduction* | 0.0769 | 0 | 3 | 0 | 17 | 0.0115 |
| ErbB signaling pathway (hsa04012)<br>*Signal Transduction* | 0.0788 | 0 | 5 | 5 | 25 | 0.0314 |
| Acute myeloid leukemia (hsa05221)<br>*Cancers* | 0.0957 | 1 | 3 | 3 | 25 | 0.0028 |
| Gap junction (hsa04540)<br>*Cell Communication* | 0.0976 | 0 | 5 | 3 | 42 | 0.0314 |

KEGG ID, Kyoto Encyclopedia of Genes and Genomes pathway identifier; P, P-value of association test in the genome-wide study; SNP, single nucleotide polymorphism; SRT P, P-value of permutation test for pathway tested

negative breast cancers. This emphasizes the importance of looking at ER-negative breast cancer separately as a unique breast cancer phenotype.

Overall, no significant signal peak was identified in this study (Figures 4, 5, 6, 7, 8). Only one SNP (rs361147) was found to achieve genome-wide significance after correction for multiple testing in the single marker analysis. However, the other loci exhibiting strong associations were interesting for reasons of biological significance, and were considered to merit further research. The associated region on 9q31.1 tagged by rs7039994 contains two known genes, *TEX10* (testis expressed sequence 10) and *INVS* (inversin). No function has been ascribed to *TEX10*. *INVS* is reported to function as a molecular switch between different Wnt signalling pathways [49] and is also pivotal in the establishment of the left-right axis. The *RHOBTB3* gene, harbouring SNP rs3777218, was identified as a putative breast cancer anti-estrogen resistance gene [50]. However, none of these single markers most strongly associated with ER-negative breast cancer could be replicated in a larger, independent sample made up of two independent studies (Table 1)

To maximize the information obtained from the GWAS scan, we conducted a permutation-based pathway analysis using the KEGG database to capture the joint actions of multiple SNPs with modest effects. In the analysis using default SRT pathway definition files comprising within-transcript SNPs, metabolic pathways involving pentose and glucuronate interconversions (hsa00040) ($P$ = 0.022) as well as starch and sucrose metabolism (hsa00500) ($P$ = 0.042) were found to be nominally significantly related to the risk of developing ER-negative breast cancer (Table 2). Estrogen-induced breast cancer cell proliferation is often accompanied by an increase in intracellular metabolic activity, resulting in a higher growth rate. The pentose phosphate pathway, which works in tight conjunction with the pentose

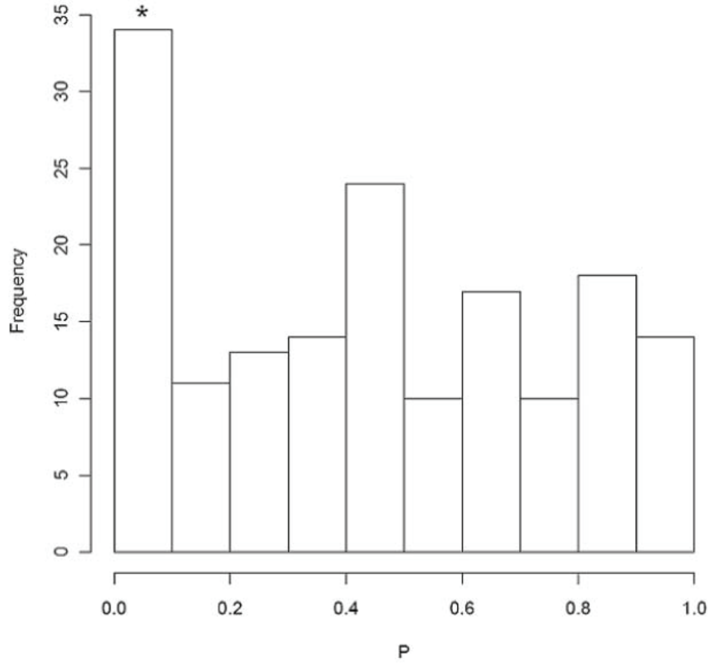


**Figure 9 Distribution of *P*-values of regulatory SNPs within KEGG cancer pathways (pathway identifiers beginning with hsa052*)**. *Global *P*-values of cancer-related regulatory SNPs with *P* < 0.05 in the genome-wide association analysis using the admixture maximum likelihood test (5,000 permutations) are 0.0028 (unadjusted), and 0.052 (with adjustments made to correct for population stratification).

and glucuronate interconversions and starch and sucrose metabolism pathways, has recently been suggested to be essential for estrogen-dependent cell proliferation [51]. Several pathways that were found to be marginally significant ($P < 0.1$) have been suggested to have potential roles in ER-negative breast cancer, namely, the TGF-beta signalling pathway [52], the renin-angiotensin system [53], and the Notch signalling pathway [54]. In addition, the insulin signalling pathway has been the focus of targeted therapy for breast cancer [55], and the purine metabolism pathway is also closely related to the pentose phosphate pathway described earlier.

Nevertheless, there is neither a precise biological definition of a pathway, nor a "standard" method to map SNPs to genes, and then genes to pathways. Pathway analyses of GWAS of common diseases have mostly based SNP-to-gene mappings on the chromosomal position of the SNP, whether it occurs within transcript of a certain gene [19,56]. However, it may be more meaningful to map SNPs that are associated with the expression of a gene to the gene. To elucidate pathways with more biological relevance, we further conducted pathway analysis based on a subset of SNPs with known regulatory functions. Recent studies have observed that whereas stronger effects overlap between different tissues, weak effects on gene regulation are tissue-specific [57,58]. Since we utilized data on gene regulation from lymphoblasts, we decided to restrict our dataset to only genes regulated on a genome-wide significant level (LOD >6). This minimized the bias of tissue-specific gene regulation, but at the same time, limited us to only a fraction of all possible SNPs genotyped within our GWAS, thus reducing the power of the analysis.

In spite of the limitations, four of the five significantly associated pathways ($P < 0.05$) in our analysis were found to be annotated as cancer pathways in KEGG (glioma (hsa05214), non-small cell lung cancer (hsa05223), pancreatic cancer (hsa05212), and prostate cancer (hsa05215) (Table 3)), hence confirming the validity of the choice of this subset of regulatory SNPs in pathway definition. In addition, a global test of the

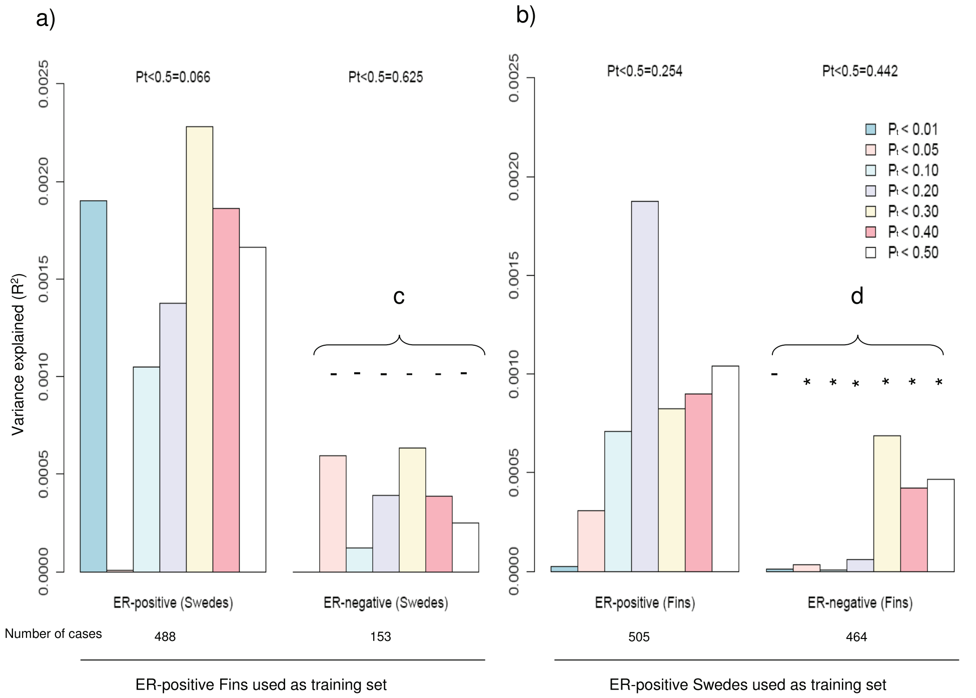


**Figure 10 Proportion of shared polygenic component between breast cancer estrogen receptor subtypes**. Proportion of shared polygenic component between ER-positive and ER-negative target samples, with respect to their corresponding ER-positive training samples. Pt denotes *P*-value cut-off in training sample. **a)** Test for association between polygenic score and disease status (ER-positive/ER-negative) in the Swedish data, when all SNPs with $P < 0.5$ in the Finnish training set were considered. **b)** Test for association between polygenic score and disease status (ER-positive/ER-negative) in the Finnish data, when all SNPs with $P < 0.5$ in the Swedish training set were considered. **c)** Significance test for difference in scores (Finnish ER-positive breast cancers derived) between Swedish ER-negative and ER-positive breast cancers, adjusted for number of non-missing genotypes. Significance codes: '- ' $0.1 < P < 1$ (that is, not significant). **d)** Significance test for difference in scores (Swedish ER-positive breast cancers derived) between Finnish ER-negative and ER-positive breast cancers, adjusted for number of non-missing genotypes. Significance codes: '*' $0.01 < P < 0.05$.

SNPs defining the cancer pathways found the aggregate effect to be approaching statistical significance ($P_{\alpha = 0.05} = 0.052$). Due to the large number of markers evaluated in a genome-wide scan, signals with small effects and modestly significant *P*-values are likely to be dismissed after the correction of multiple testing. The implementation of a pathway analysis thus serves as a complementation between a hypothesis-driven (prior knowledge of biological pathways) and a hypothesis-free (genome-wide scan) approach to highlight certain markers, such as those found in the cancer pathways, worthy of further study that would not have been examined otherwise. The lack of a concordance between the results of pathway analyses using two different SNP-to-gene mapping approaches emphasizes the need to put in more consideration in choosing appropriate pathway definitions. An excess of small *P*-values found for SNPs associated with gene expression involved in cancer-related pathways suggests that the SNP-gene mapping via association with gene expression approach is superior to the SNP-gene mapping by location within a transcript approach, and should be explored in greater detail.

Limitations of this study include an overall lack of statistical power, especially for the single marker analysis, and the existence of further heterogeneity among ER-negative tumours. Although genome-wide pathway-based analysis is an interesting approach, a main limitation is that the associations observed in this study are only nominally significant, and would not be significant after correction for multiple testing. However, as many pathways have SNPs in common with other pathways, the stringent significance thresholds of traditional multiple testing correction methods are potentially over-conservative. There is also indirect evidence that corroborates our pathway findings. Gene expression studies have found pathways related to the renin-angiotensin system and focal adhesion to be significantly associated with prognosis of breast cancer [59]. Others have also reported pathways highlighted in our study, which are involved in pentose and glucuronate interconversions, gap junction, TGF-beta signalling, rennin-angiotensin system, B cell receptor signalling, Fc epsilon RI signalling, VEGF signalling, ErbB signalling, and focal adhesion, to be significantly associated with the breast cancer phenotype [59,60]. Although replication of the pathway results in independent studies would be needed to confirm the associations, the substantial additional sample collection and genotyping required are beyond the scope of this publication.

Although breast cancer has been classified into ER-positive and ER-negative breast cancers, and these two breast cancer subtypes have been documented to show different gene expression patterns, GWAS scans on breast cancer have always been performed on either overall breast cancer (ER-positive, ER-negative and unknown) or ER-positive breast cancer specific risks. In this study, we found evidence to suggest that ER-negative breast cancers only share a fraction of the polygenic component of the disease with ER-positive breast cancers, implying that ER-negative breast cancer should be examined as a distinct breast cancer phenotype. Although the difference between the polygenic components of ER-positive and ER-negative breast cancers was found only to be significant in the Finnish training samples, we observed similar differences for all seven *P*-value thresholds in the Swedish training samples. However, due to the smaller number of Swedish ER-negative cases (*N* = 153, approximately 33% of Finnish ER-negative cases), we had less power to detect significant heterogeneity between the two subtypes in the Swedish target samples.

## Conclusions

Given the clinical importance of the ER-negative phenotype and the likelihood that the relative genetic effect sizes are small, greater sample sizes and further studies are required to further the knowledge on ER-negative breast cancers. Identification of factors for a predisposition to ER-negative tumours opens the way for understanding the underlying etiology of the disease, and may ultimately result in improvements in prevention, early detection and specific treatment for this tumour subtype. We used a novel approach to pathway analysis, showing that established cancer pathways could be regulated by common variants associated to ER-negative breast cancer. We also provided molecular genetic evidence which suggests that ER-negative breast cancer is a distinct breast cancer subtype that merits independent analyses. In view of the biological relevance of the pathways identified, a genome-wide pathway approach deserves merit, and has good potential in pointing out directions for future research for ER-negative breast cancers.

## Additional material

**Additional file 1: Supplementary Methods**. Full methods accompanying this manuscript.

**Additional file 2: Supplementary figures**. Supplementary Figure 1. Scree plot of log-transformed Eigen values. Vertical dashed lines indicate three and five PCs taken to correct for population stratification within the Swedish and Finnish populations respectively. Supplementary Figure 2. Quantile-quantile plot for 285,984 SNP trend tests, adjusted for population stratification using three principal components (Swedish subjects only). Genomic control inflation factor (λ) = 1.0140. Supplementary Figure 3. Quantile-quantile plot for 285,984 SNP trend tests, adjusted for population stratification using five principal components (Finnish subjects only). Genomic control inflation factor (λ) = 1.0137. Supplementary Figure 4. Quantile-quantile plot for 285,984 SNP trend tests, adjusted for population stratification (combined analysis of Swedish and Finnish subjects). Genomic control inflation factor (λ) = 1.0218.

**Additional file 3: Supplementary tables**. Table S1. Association analysis results of reclustered SNPs. Table S2. Association results of top hits in the combined analysis, with corresponding MAF, ORs and P values within the Swedish and Finnish populations. * denotes the five SNPs selected for validation in SEARCH and RBCS. Table S3. Power to detect single marker effects in genome-wide association study.

### Abbreviations

Λ: genomic control inflation factor; AML: admixture maximum likelihood; EIRA: Epidemiological Investigation of Rheumatoid Arthritis; ER: estrogen receptor; GWAS: genome-wide association study; HWE: Hardy-Weinberg equilibrium; KEGG: Kyoto Encyclopedia of Genes and Genomes; LD: linkage disequilibrium; MAF: minor allele frequency; NCBI: National Center for Biotechnology Information; PC: principal component; PCA: principal component analysis; RBCS: Rotterdam Breast Cancer Study; SEARCH: Study of Epidemiology and Risk factors in Cancer Heredity; SNP: single nucleotide polymorphism; SRT: SNP ratio test.

### Acknowledgements

The Swedish study was supported by the National Institutes of Health (RO1 CA58427 to PH and JLiu), the Agency for Science, Technology and Research (A*STAR; Singapore), the Nordic Cancer Society, Märit and Hans Rausing's Initiative against Breast Cancer, and the Cancer Risk Prediction Center (CRisP), a LinneusCentre (Contract ID 70867902) financed by the Swedish researchCouncil. J Li is a recipient of an A*STAR Graduate Scholarship (Overseas) award. KH was supported by the Swedish Research Council (523-2006-972). KC was supported by the Swedish Cancer Society (5128-B07-01PAF). The Helsinki study was supported by the Nordic Cancer Society, the Helsinki University Central Hospital Research Fund, the Academy of Finland (132473), the Finnish Cancer Society and the Sigrid Juselius Foundation. We thank Dr. Kirsimari Aaltonen and RN Hanna Jäntti for their help with the patient data. The SEARCH study was funded by Cancer Research UK. DFE is a Principal Research Fellow of Cancer Research UK. We thank the SEARCH team and Eastern Cancer Registry and Information Centre (ECRIC) for recruitment of the SEARCH cases. The RBCS was supported by the Dutch Cancer Society (grant DDHK 2004-3124). We acknowledge the clinicians from the Rotterdam Family Cancer Clinic who were involved in collecting the RBCS samples: C. Seynaeve, J. Klijn, J. Collee, and R. Oldenburg.

### Author details

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, P.O. Box 281, Stockholm 17177, Sweden. [2]Human Genetics, Genome Institute of Singapore, 60 Biopolis St, Singapore 138672, Singapore. [3]Department of Biosciences and Nutrition, Karolinska Institutet, Hälsovägen 7-9, Novum, SE-141 81, Huddinge, Sweden. [4]Department of Obstetrics and Gynecology, Helsinki University Central Hospital, P.O. Box 700, 00029 HUS, Helsinki, Finland. [5]Department of Clinical Genetics, Helsinki University Central Hospital, Haartmanink 2 B, 00029 HUS, Helsinki, Finland. [6]Department of Oncology, Helsinki University Central Hospital, P.O. Box 180, 00029 HUS, Helsinki, Finland. [7]Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Wort's Causeway, Cambridge CB1 8RN, UK. [8]Department of Oncology, Strangeways Research Laboratory, University of Cambridge, Wort's Causeway, Cambridge CB1 8RN, UK. [9]Department of Medical Oncology, Rotterdam Family Cancer Clinic, Erasmus University Medical Center, Daniel den Hoed Cancer Center, Groene Hilledijk 301, 3075 EA Rotterdam, Netherlands. [10]Department of Medical Oncology, Erasmus University Medical Center, Josephine Nefkens Institute, Dr. Molenwaterplein 50, 3015 GE Rotterdam, The Netherlands. [11]Department of Clinical Genetics, Rotterdam Family Cancer Clinic, Erasmus University Medical Center, Dr. Molenwaterplein 50, 3015 GE Rotterdam, Netherlands. [12]Institute of Environmental Medicine, Karolinska Institutet, P.O. Box 281, Stockholm 17177, Sweden. [13]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, P.O. Box 20, FI-00014, Finland. [14]Public Health Genomics Unit, National Institute for Health and Welfare, P.O. Box 30, FI-00271 Helsinki, Finland. [15]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [16]Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [17]Clinical Research Centre, Karolinska Institute, Karolinska University Hospital Huddinge, SE-141 86, Huddinge, Sweden. [18]Department of Medical Genetics, University of Helsinki, Haartman Institute, P.O. Box 21 (Haartmaninkatu 3), FI-00014, Finland. [19]Folkhälsan Institute of Genetics, Folkhälsan Research Center; University of Helsinki, Haartmaninkatu 8, Biomedicum 1, P.O. Box 63, FI-00014, Finland.

### Authors' contributions

JLi, KH, HN, JLiu, KC, and PH conceived and designed the experiments. JLi, KH, HD, UH, TH, AI, HQL, GHKT, AT and GR analyzed the data. KA, CB, PDPP, AMD, DA, MJH, AH, RAO, LA, AP, LPP, JK, MD, DFE, HN, JLiu, KC and PH contributed reagents/materials/analysis tools. JLi, KH, HD, GR, UH, TH, KA, CB, PDPP, AMD, DA, MJH, AH, RAO, LA, AP, LPP, AI, HQL, GHKT, AT, JK, MD, DFE, HN, JLiu, KC and PH wrote the paper.

### Competing interests

The authors declare that they have no competing interests.

Received: 5 August 2010 Revised: 6 October 2010
Accepted: 9 November 2010 Published: 9 November 2010

### References

1. Perou CM, Sorlie 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, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
2. Sorlie 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, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98**:10869-10874.
3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland.** *N Engl J Med* 2000, **343**:78-85.
4. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, *et al*: **A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1).** *Nat Genet* 2009, **41**:579-584.
5. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447**:1087-1093.
6. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S, Lissowska J, Brinton L, Peplonska B, Southey MC, Hopper JL, McCredie MR, Giles GG, Fletcher O, Johnson N, dos Santos Silva I, Gibson L, Bojesen SE, Nordestgaard BG, Axelsson CK, Torres D, Hamann U, Justenhoven C, Brauch H, Chang-Claude J, Kropp S, *et al*: **A common coding variant in CASP8 is associated with breast cancer risk.** *Nat Genet* 2007, **39**:352-358.
7. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39**:870-874.
8. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, *et al*: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2007, **39**:865-869.
9. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, *et al*: **Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer.** *Nat Genet* 2008, **40**:703-706.
10. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, Eccles D, Evans DG, Fletcher O, Johnson N, dos Santos Silva I, Peto J, Stratton MR, Rahman N, Jacobs K, Prentice R, Anderson GL, Rajkovic A, Curb JD, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, *et al*: **Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2.** *Nat Genet* 2009, **41**:585-590.
11. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, Gu K, Fair AM, Cai Q, Lu W, Shu XO: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1.** *Nat Genet* 2009, **41**:324-328.
12. Thomas D: **Gene-environment-wide association studies: emerging approaches.** *Nat Rev Genet* **11**:259-272.
13. Pedroso I: **Gaining a pathway insight into genetic association data.** *Methods Mol Biol* **628**:373-382.
14. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR: **Pathway and network-based analysis of genome-wide association studies in multiple sclerosis.** *Hum Mol Genet* 2009, **18**:2078-2090.
15. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC: **Using genome-wide pathway analysis to unravel the etiology of complex diseases.** *Genet Epidemiol* 2009, **33**:419-431.
16. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M: **Gene and pathway-based second-wave analysis of genome-wide association studies.** *Eur J Hum Genet* **18**:111-117.
17. Ritchie MD: **Using prior knowledge and genome-wide association to identify pathways involved in multiple sclerosis.** *Genome Med* 2009, **1**:65.
18. Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies.** *Am J Hum Genet* 2007.

19. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A: **The SNP ratio test: pathway analysis of genome-wide association datasets.** *Bioinformatics* 2009, **25**:2762-2763.
20. Guo YF, Li J, Chen Y, Zhang LS, Deng HW: **A new permutation strategy of pathway-based approach for genome-wide association study.** *BMC Bioinformatics* 2009, **10**:429.
21. Garcia-Closas M, Chanock S: **Genetic susceptibility loci for breast cancer by estrogen receptor status.** *Clin Cancer Res* 2008, **14**:8000-8009.
22. Mavaddat N, Pharoah PD, Blows F, Driver KE, Provenzano E, Thompson D, Macinnis RJ, Shah M, Easton DF, Antoniou AC: **Familial relative risks for breast cancer by pathological subtype: a population-based cohort study.** *Breast Cancer Res* **12**:R10.
23. Lesueur F, Pharoah PD, Laing S, Ahmed S, Jordan C, Smith PL, Luben R, Wareham NJ, Easton DF, Dunning AM, Ponder BA: **Allelic association of the human homologue of the mouse modifier Ptprj with breast cancer.** *Hum Mol Genet* 2005, **14**:2349-2356.
24. Magnusson C, Baron J, Persson I, Wolk A, Bergstrom R, Trichopoulos D, Adami HO: **Body size in different periods of life and breast cancer risk in post-menopausal women.** *Int J Cancer* 1998, **76**:29-34.
25. Rosenberg LU, Einarsdottir K, Friman EI, Wedren S, Dickman PW, Hall P, Magnusson C: **Risk factors for hormone receptor-defined breast cancer in postmenopausal women.** *Cancer Epidemiol Biomarkers Prev* 2006, **15**:2482-2488.
26. Li J, Humphreys K, Heikkinen T, Aittomaki K, Blomqvist C, Pharoah PD, Dunning AM, Ahmed S, Hooning MJ, Martens JW, van den Ouweland AM, Alfredsson L, Palotie A, Peltonen-Palotie L, Irwanto A, Low HQ, Teoh GH, Thalamuthu A, Easton DF, Nevanlinna H, Liu J, Czene K, Hall P: **A combined analysis of genome-wide association studies in breast cancer.** *Breast Cancer Res Treat* .
27. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK: **TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study.** *N Engl J Med* 2007, **357**:1199-1209.
28. Syrjakoski K, Vahteristo P, Eerola H, Tamminen A, Kivinummi K, Sarantaus L, Holli K, Blomqvist C, Kallioniemi OP, Kainu T, Nevanlinna H: **Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients.** *J Natl Cancer Inst* 2000, **92**:1529-1531.
29. Kilpivaara O, Bartkova J, Eerola H, Syrjakoski K, Vahteristo P, Lukas J, Blomqvist C, Holli K, Heikkila P, Sauter G, Kallioniemi OP, Bartek J, Nevanlinna H: **Correlation of CHEK2 protein expression and c.1100delC mutation status with tumor characteristics among unselected breast cancer patients.** *Int J Cancer* 2005, **113**:575-580.
30. Fagerholm R, Hofstetter B, Tommiska J, Aaltonen K, Vrtel R, Syrjakoski K, Kallioniemi A, Kilpivaara O, Mannermaa A, Kosma VM, Uusitupa M, Eskelinen M, Kataja V, Aittomaki K, von Smitten K, Heikkila P, Lukas J, Holli K, Bartkova J, Blomqvist C, Bartek J, Nevanlinna H: **NAD(P)H:quinone oxidoreductase 1 NQO1*2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer.** *Nat Genet* 2008, **40**:844-853.
31. Eerola H, Blomqvist C, Pukkala E, Pyrhonen S, Nevanlinna H: **Familial breast cancer in southern Finland: how prevalent are breast cancer families and can we trust the family history reported by patients?** *Eur J Cancer* 2000, **36**:1143-1148.
32. Bilguvar K, Yasuno K, Niemela M, Ruigrok YM, von Und Zu Fraunberg M, van Duijn CM, van den Berg LH, Mane S, Mason CE, Choi M, Gaal E, Bayri Y, Kolb L, Arlier Z, Ravuri S, Ronkainen A, Tajima A, Laakso A, Hata A, Kasuya H, Koivisto T, Rinne J, Ohman J, Breteler MM, Wijmenga C, State MW, Rinkel GJ, Hernesniemi J, Jaaskelainen JE, Palotie A, *et al*: **Susceptibility loci for intracranial aneurysm in European and Japanese populations.** *Nat Genet* 2008, **40**:1472-1477.
33. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BW, Janssens AC, Wilson JF, Spector T, Martin NG, Pedersen NL, Kyvik KO, Kaprio J, Hofman A, Freimer NB, Jarvelin MR, Gyllensten U, Campbell H, Rudan I, Johansson A, Marroni F, Hayward C, Vitart V, Jonasson I, Pattaro C, Wright A, Hastie N, Pichler I, Hicks AA, *et al*: **Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts.** *Nat Genet* 2009, **41**:47-55.
34. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Jarvelin MR, Freimer NB, Peltonen L: **Genome-wide association analysis of metabolic traits in a birth cohort from a founder population.** *Nat Genet* 2009, **41**:35-46.
35. Leu M, Humphreys K, Surakka I, Rehnberg E, Muilu J, Rosenström P, Almgren P, Jääskeläinen J, Lifton RP, Kyvik KO, Kaprio J, Pedersen NL, Palotie A, Hall P, Grönberg H, Groop L, Peltonen L, Palmgren J, Ripatti S: **NordicDB: A Nordic pool and portal for genome-wide control data.** *Eur J Hum Genet* 2010.
36. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH: **A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.** *Science* 2006, **314**:1461-1463.
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559-575.
38. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904-909.
39. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.
40. **mRNA by SNP Browser v 1.0.1.** [http://www.sph.umich.edu/csg/liang/asthma/].
41. Tyrer J, Pharoah PD, Easton DF: **The admixture maximum likelihood test: a novel experiment-wise test of association between disease and multiple SNPs.** *Genet Epidemiol* 2006, **30**:636-643.
42. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P: **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.** *Nature* 2009, **460**:748-752.
43. R Development Core Team: **R: A Language and Environment for Statistical Computing.** Vienna, Austria: R Foundation for Statistical Computing; 2007.
44. **QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies.** [http://hydra.usc.edu/gxe].
45. **Qlikview.** [http://www.qliktech.com].
46. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**:263-265.
47. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ: **LocusZoom: regional visualization of genome-wide association scan results.** *Bioinformatics* 2010, **26**:2336-2337.
48. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO: **A genome-wide association study of global gene expression.** *Nat Genet* 2007, **39**:1202-1207.
49. Simons M, Gloy J, Ganner A, Bullerkotte A, Bashkurov M, Kronig C, Schermer B, Benzing T, Cabello OA, Jenny A, Mlodzik M, Polok B, Driever W, Obara T, Walz G: **Inversin, the gene product mutated in nephronophthisis type II, functions as a molecular switch between Wnt signaling pathways.** *Nat Genet* 2005, **37**:537-543.
50. van Agthoven T, Veldscholte J, Smid M, van Agthoven TL, Vreede L, Broertjes M, de Vries I, de Jong D, Sarwari R, Dorssers LC: **Functional identification of genes causing estrogen independence of human breast cancer cells.** *Breast Cancer Res Treat* 2009, **114**:23-30.
51. Forbes NS, Meadows AL, Clark DS, Blanch HW: **Estradiol stimulates the biosynthetic pathways of breast cancer cells: detection by metabolic flux analysis.** *Metab Eng* 2006, **8**:639-652.
52. Biswas S, Guix M, Rinehart C, Dugger TC, Chytil A, Moses HL, Freeman ML, Arteaga CL: **Inhibition of TGF-beta with neutralizing antibodies prevents radiation-induced acceleration of metastatic cancer progression.** *J Clin Invest* 2007, **117**:1305-1313.
53. Herr D, Rodewald M, Fraser HM, Hack G, Konrad R, Kreienberg R, Wulff C: **Potential role of Renin-Angiotensin-system for tumor angiogenesis in receptor negative breast cancer.** *Gynecol Oncol* 2008, **109**:418-425.
54. Dontu G, Jackson KW, McNicholas E, Kawamura MJ, Abdallah WM, Wicha MS: **Role of Notch signaling in cell-fate determination of human mammary stem/progenitor cells.** *Breast Cancer Res* 2004, **6**:R605-615.
55. Zeng X, Yee D: **Insulin-like growth factors and breast cancer therapy.** *Adv Exp Med Biol* 2007, **608**:101-112.

56. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, Kraft P, Hunter DJ, Chanock SJ, Rosenberg PS, Chatterjee N: Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Res* **70**:4453-4459.
57. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE: Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 2009, **325**:1246-1250.
58. Kwan T, Grundberg E, Koka V, Ge B, Lam KC, Dias C, Kindmark A, Mallmin H, Ljunggren O, Rivadeneira F, Estrada K, van Meurs JB, Uitterlinden A, Karlsson M, Ohlsson C, Mellstrom D, Nilsson O, Pastinen T, Majewski J: Tissue effect on genetic control of transcript isoform variation. *PLoS Genet* 2009, **5**:e1000608.
59. Ma S, Kosorok MR: Detection of gene pathways with predictive power for breast cancer prognosis. *BMC Bioinformatics* **11**:1.
60. Gohlke JM, Thomas R, Zhang Y, Rosenstein MC, Davis AP, Murphy C, Becker KG, Mattingly CJ, Portier CJ: Genetic and environmental pathways to complex diseases. *BMC Syst Biol* 2009, **3**:46.

doi:10.1186/bcr2772
**Cite this article as:** Li *et al.*: A genome-wide association scan on estrogen receptor-negative breast cancer. *Breast Cancer Research* 2010 **12**:R93.

# III

Li *et al. Breast Cancer Research* 2010, **12**:R19
http://breast-cancer-research.com/content/12/2/R19

Breast Cancer RESEARCH

RESEARCH ARTICLE Open Access

# Genetic variation in the estrogen metabolic pathway and mammographic density as an intermediate phenotype of breast cancer

Jingmei Li[1,2*], Louise Eriksson[1], Keith Humphreys[1], Kamila Czene[1], Jianjun Liu[2], Rulla M Tamimi[3,4], Sara Lindström[5], David J Hunter[3], Celine M Vachon[6], Fergus J Couch[7], Christopher G Scott[6], Pagona Lagiou[4,8], Per Hall[1]

## Abstract

**Introduction:** Several studies have examined the effect of genetic variants in genes involved in the estrogen metabolic pathway on mammographic density, but the number of loci studied and the sample sizes evaluated have been small and pathways have not been evaluated comprehensively. In this study, we evaluate the association between mammographic density and genetic variants of the estrogen metabolic pathway.

**Methods:** A total of 239 SNPs in 34 estrogen metabolic genes were studied in 1,731 Swedish women who participated in a breast cancer case-control study, of which 891 were cases and 840 were controls. Film mammograms of the medio-lateral oblique view were digitalized and the software Cumulus was used for computer-assisted semi-automated thresholding of mammographic density. Generalized linear models controlling for possible confounders were used to evaluate the effects of SNPs on mammographic density. Results found to be nominally significant were examined in two independent populations. The admixture maximum likelihood-based global test was performed to evaluate the cumulative effect from multiple SNPs within the whole metabolic pathway and three subpathways for androgen synthesis, androgen-to-estrogen conversion and estrogen removal.

**Results:** Genetic variants of genes involved in estrogen metabolism exhibited no appreciable effect on mammographic density. None of the nominally significant findings were validated. In addition, global analyses on the overall estrogen metabolic pathway and its subpathways did not yield statistically significant results.

**Conclusions:** Overall, there is no conclusive evidence that genetic variants in genes involved in the estrogen metabolic pathway are associated with mammographic density in postmenopausal women.

## Introduction

Mammographic breast density is one of the strongest risk factors for breast cancer. Several studies have shown that women with extensive dense tissue are at two to six times higher risk of developing the disease than women of similar age with lower mammographic density [1,2]. A strong genetic basis has been suggested for mammographic density [3]. Twin studies have estimated the heritability of this trait to be between 60 and 67% [4]. Evidence for a genetic influence also comes from other studies on family history, familial aggregation and segregation analyses [5,6].

Mammographic density is strongly correlated with hormone exposure profiles of women [7]. Several hormonal risk factors for breast cancer have been found to influence mammographic density in a similar fashion to their respective associations with risk for the disease [8]. For example, a strong inverse relationship has been observed between parity on mammographic density [9]. In addition, hormone replacement therapy (HRT) users and women who have a late first-born child or late menopause have higher breast densities on average [9]. In view of evidence suggesting an association between mammographic density and hormone-related factors, and the fact that estrogen is a strong risk factor for postmenopausal breast cancer, efforts have been made to identify underlying genetic determinants of

* Correspondence: Jingmei.Li@ki.se
[1]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Box 281, 171 77 Stockholm, Sweden

mammographic density within pathways related to steroid hormone biosynthesis and metabolism [10-13]. Such endeavors assume mammographic density to be an intermediate phenotype for breast cancer. Several genes involved in hormone-related pathways - such as HSD3B1 [5,14], COMT [11,14] and ESR1 [15] - have been suggested to be associated with mammographic breast density. Findings are inconsistent, however, and only few candidate genes have been studied at a time.

We recently reported the results of a study evaluating a total of 239 SNPs in 34 estrogen metabolic genes in 1,596 breast cancer cases and 1,730 population controls from Sweden, of which the outcome variable was breast cancer (Low *et al*., manuscript submitted). No significant SNP association was evident after correction for multiple testing, but pathway-based global tests revealed significant association evidence for the overall estrogen metabolic pathway ($P = 0.034$) and, in particular, the androgen-to-estrogen conversion subpathway ($P = 0.007$). In the present study, we comprehensively examine genetic variation in the estrogen metabolic pathway with mammographic density. The number of SNPs and genes studied provides the most extensive coverage to date with respect to studying mammographic breast density.

## Materials and methods

### Study subjects

The subjects included in the current study are drawn from a population-based case-control study of postmenopausal breast cancer in women born in Sweden aged 50 to 74 years at the time of enrollment, which was between 1 October 1993 and 31 March 1995. Controls were randomly selected from the Swedish Total Population Register and were frequency matched to the expected age distribution of the cases. Details on data collection and subjects have been described previously [16]. The final study group with both mammographic density and genotype data included 891 breast cancer cases and 840 controls. Although all women were postmenopausal at the time of recruitment to the parent study, a subset of the women (43/1,731) was premenopausal in reference to the date of mammogram.

Approval of the study was given by the ethical review board at the Karolinska Institutet (Stockholm, Sweden) and six other ethical review boards in the respective regions in which the subjects were based, and informed consent was obtained from each participant.

Validation of SNPs with significant associations was performed using mammographic density data from two other studies.

### Mammographic density data

The process of collecting mammographic density data in this study has been described previously [17]. Film mammograms of the medio-lateral oblique view were digitized using an Array 2905HD Laser Film Digitizer (Array Corporation, Tokyo, Japan), which covers a range of 0 to 4.7 optical density. For controls, the breast side was randomized. For cases, the side contralateral to the tumor was used. The density resolution was set at 12-bit spatial resolution. The Cumulus software used for the computer-assisted measurement was developed at the University of Toronto [18]. For each image, a trained observer (LE) set the appropriate gray-scale threshold levels defining the edge of the breast and distinguishing dense from nondense tissue. The software calculated the total number of pixels within the entire region of interest and within the region identified as dense. These values were used to calculate the percentage of the breast area that is dense. A random 10% of the images were included as replicates to assess the intra-observer reliability, which was high with a Spearman rank correlation coefficient of 0.95.

### Gene and SNP selection

The process of gene and SNP selection has been described in detail by Low *et al.* (manuscript submitted). A total of 1,007 SNPs were selected from 35 genes and their 30 kb flanking sequences that code the enzymes involved in estradiol or estrone metabolism and are expressed in the breast. These SNPs were genotyped in 92 Swedish control samples to assess linkage disequilibrium patterns, to select tagging SNPs (tagSNPs) and to evaluate their coverage.

Haplotypes were reconstructed using the partition-ligation-expectation-maximization algorithm [19] implemented in the *tagSNPs* program [20]. A subset of tagSNPs were selected based on the $R^2$ coefficient, which quantifies how well the tagSNP haplotypes predict the genotype or the number of copies of haplotypes an individual carries. The performance of tagSNPs in capturing unobserved SNPs within the genes was evaluated using a SNP-dropping analysis. In brief, each of the genotyped SNPs was dropped in turn and then tagSNPs were selected from the remaining SNPs so that their haplotypes predicted the remaining SNPs with an $R^2$ value of 0.85. In total, 312 tagSNPs from the 35 genes were selected for genotyping.

Figure 1 delineates the processes and genes involved in the androgen synthesis, androgen-to-estrogen conversion and estrogen removal subpathways. The lists of SNPs corresponding to each subpathway are summarized in Tables S1 to S3 in Additional file 1.

### DNA extraction and genotyping

DNA was extracted from 4 ml whole blood using the QIAamp DNA Blood Maxi Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions

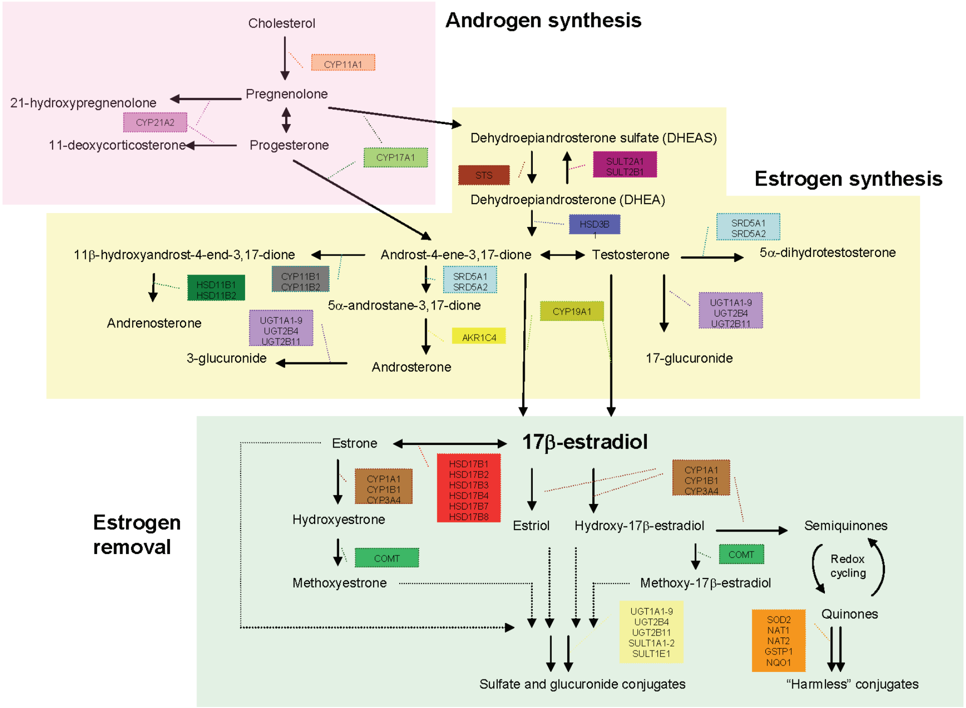


**Figure 1 Subdivision of the estrogen metabolic pathway**. The 34 metabolic genes analyzed in the present study are involved in different steps of the estrogen metabolism. The genes are divided into the three groups involved in androgen synthesis, estrogen synthesis and estrogen removal for subpathway-based association analysis.

and nonmalignant cells in paraffin-embedded tissue using a standard phenol/chloroform/isoamyl alcohol protocol. Genotyping was performed using the primer extension-based assay from Sequenom (San Diego, CA, USA) according to the manufacturers' instructions. DNA samples were randomly assigned to the plates carrying positive and negative controls, and all genotyping results were generated and checked by laboratory staff unaware of the case-control status. SNPs with a call rate <85%, minor allele frequency <1% or out of Hardy-Weinberg equilibrium ($P$ < 0.05/312) were excluded from further analysis. The genotype concordance was >99%, suggesting high genotyping accuracy. Overall, 239 tagSNPs from the 34 genes were successfully genotyped and used in statistical analysis.

### Statistical analysis

Linear regression models were fitted, treating percentage density as an outcome. Models were adjusted for age, body mass index, menopausal status and HRT. Age was coded as 0, 1 and 2 for women <50 years, between 50 and 60 years, and >60 years of age, respectively. The body mass index was treated as a continuous variable. Menopausal status was determined from the time difference between the date of menopause and the date on which the mammogram was taken. HRT was considered a categorical variable made up of three groups: never users, past users and current users. The mammographic density measurements were transformed by the power of 0.3, yielding an approximately normal distribution. The genotypes were coded 0, 1 and 2 and treated as continuous variables.

A likelihood ratio test was performed for each SNP. Normal quantile-quantile plots were used to examine the distributions of the -log$_{10}$-transformed $P$ values. To assess whether the SNPs associated with breast cancer risk are the same SNPs as those associated with mammographic density, we used the Spearman's rank correlation test, evaluating the relationship between odds ratios corresponding to SNP effects on breast cancer risk and the regression coefficients of SNP effects on percentage density. The admixture maximum likelihood-based global test [21] was performed to evaluate the cumulative effect on mammographic density from multiple SNPs within the whole metabolic pathway and three subpathways for androgen synthesis, androgen-to-estrogen conversion and estrogen removal. Affection status for the admixture maximum likelihood analysis was defined by taking the lowest quantile of all percentage density measurements as controls and the highest quantile as cases. *P* values of the admixture maximum likelihood test were obtained via 5,000 permutations. Software R (v2.8.0) [22] and admixture maximum likelihood [21] were used for data management, quality control and statistical analyses.

### Validation of significantly associated SNPs

SNP associations with mammographic density were validated in 1,590 women genotyped with the Illumina HumanHap500 as part of the Cancer Genetic Markers of Susceptibility Project (CGEMS) [23]. The CGEMS project is a National Cancer Institute initiative to conduct genome-wide association studies to identify genes involved in breast cancer and prostate cancer. The initial CGEMS breast cancer scan was designed and funded to study the main effect of SNP variants on breast cancer risk in postmenopausal women, and has been completed [24]. Briefly, the first stage of the project involved a whole genome scan of 1,145 invasive postmenopausal breast cancer cases and 1,142 matched controls from the Nurses' Health Study nested case-control study [24]. The Nurses' Health Study was initiated in 1976, when 121,700 US registered nurses aged 30 to 55 returned an initial questionnaire [25]. During 1989 and 1990, blood samples were collected from 32,826 women [26]. For 1,590 of these women - of which 806 were breast cancer cases and 784 were healthy controls - we also had mammographic density measurements.

We collected mammograms as close as possible to the date of blood collection (1989 to 1990). To assess mammographic density, the craniocaudal (CC) views of both breasts were digitized at 261 μm/pixel with a Lumysis 85 laser film scanner, which covers a range of 0 to 4.0 optical density. The software for computer-assisted thresholding was developed at the University of Toronto [18]. We used the average percentage density of both breasts for this analysis. This collection has been described in detail in a previous publication [27]. SNPs not available on the Illumina HumanHap550 panel were imputed using MACH [28] based on HapMap Phase II (release 21a). For the analysis of imputed data, the ProbABEL package from the ABEL set of programs was used [29]. Percentage density was transformed by the power of 0.3 to be consistent with the parent study. This study was approved by the Committee on the Use of Human Subjects in Research at Brigham and Women's Hospital.

The second validation population consisted of a set of controls from an ongoing breast cancer case-control study at the Mayo Clinic. Briefly, the Mayo Clinic Breast Cancer Study is an Institutional Review Board-approved, clinic-based, case-control study initiated in February 2001 at Mayo Clinic, Rochester, MN, USA. The study design has been presented previously [30,31]. Clinic attendance formed the sampling frame for Mayo Clinic cases and controls. Consecutive cases were women aged 18 years or over with histologically confirmed primary invasive breast carcinoma and recruited within 6 months of the date of diagnosis. Cases lived in the six-state region that defines Mayo Clinic's primary service population (Minnesota, Iowa, Wisconsin, Illinois, North Dakota, and South Dakota). Controls without prior history of cancer (other than nonmelanoma skin cancer) were frequency matched on age (5-year age category), race and six-state region of residence to cases. Controls were recruited from the outpatient practice of the Divisions of General Internal Medicine and Primary Care Internal Medicine at Mayo Clinic, where they were seen for routine medical examinations.

The present analysis genotyped Caucasian controls (99% of study participants) enrolled through September 2007, who had mammograms available, representing 995 total controls (76% of total possible controls), of which 783 were postmenopausal. Screening mammograms were ascertained close to the enrollment date and the left CC view was digitized on an Array 2905HD Laser Film Digitizer, which covers a range of 0 to 4.7 optical density. Percentage mammographic density was estimated by an expert reader [32] on the left CC view, using the same Cumulus software described above [33]. Genotyping was carried out using TaqMan (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions, using 10 to 20 ng DNA. Primers and probes were Assay-by Design (Applied Biosystems). Following PCR amplification, end reactions are read on the ABI Prism 7900 ht using Sequence Detection Software (Applied Biosystems). SNP associations were examined only in the 783 postmenopausal controls, to be comparable with the two other populations. The percentage density was transformed by the power of 0.3 to be consistent with the parent study.

## Results

Our dataset consisted of 1,731 postmenopausal women, of which 981 were breast cancer cases and 840 were controls (Table 1). Cases and controls differed significantly in age at first birth ($P = 0.0126$), parity ($P < 0.0001$), family history of breast cancer ($P = 0.0002$) and

**Table 1 Selected characteristics of subjects**

| | Breast cancer cases (n = 891) | | Breast cancer controls (n = 840) | | |
|---|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** | ***P* value** |
| Age (years) | 63.0 | 6.3 | 63.0 | 6.3 | 0.9045 |
| Height (cm) | 164.1 | 5.7 | 163.6 | 5.5 | 0.0766 |
| Weight (kg) | 68.9 | 110 | 68.8 | 11.6 | 0.8153 |
| Body mass index | 25.6 | 3.9 | 25.6 | 4.1 | 0.8420 |
| Age at menarche (years) | 13.6 | 1.4 | 13.6 | 1.5 | 0.6090 |
| Age at first birth (years) | 25.4 | 50 | 24.8 | 4.7 | 0.0126 |
| Parity | 1.9 | 1.2 | 2.2 | 1.3 | 0.0000 |
| Age at menopause (years) | 50.3 | 3.6 | 50.1 | 3.9 | 0.1223 |
| HRT (% ever use) | 0.53 | | 0.50 | | 0.2523 |
| Family history (%) | 0.15 | | 0.09 | | 0.0002 |
| Percent density | 16.7 | 14.3 | 14.6 | 14.0 | 0.0017 |

Means and standard deviations (SD) are given for continuous measures, proportions for other variables. *P* values based on the Welch *t*test for independent samples. HRT, hormone replacement therapy.

percentage density (*P* = 0.0017). Cases were found to have higher percentage density (mean ± standard deviation: 16.7 ± 14.3) than controls (14.6 ± 14.0). No significant difference was found for age, height, weight, body mass index, age at menarche, age at menopause or HRT usage.

Table S4 in Additional file 2 shows a list of 34 genes involved in the estrogen metabolic pathway and the corresponding number of SNPs examined for each gene. References are given for genes that have been examined in other studies for an association with mammographic density. Of the 239 SNPs analyzed, 11 SNPs were found to be significant at the 5% level (Table 2) - of which the smallest *P* value was 0.0019. Among six tagSNPs selected for the gene CYP11A1, five were found to be significant in the same direction. The associations in the single SNP analysis were moderate and would not survive correction for multiple SNP testing. In addition, the single-SNP *P* values showed no clear deviation from the null distribution, representing no association between SNPs and percentage density (Figure 2; see also Tables S1 to S3 in Additional file 1). None of the SNPs found to be nominally significant in our dataset were found to be significant in the CGEMS validation set (see Table S5 in Additional file 3). A second, independent validation carried out on the most significantly associated SNP (rs11638442) located within the CYP11A1 gene in 783 postmenopausal women with mammograms in the Mayo Clinic Breast Cancer Study yielded a *P* value of 0.88 (regression coefficient = -0.000507, 95% confidence interval = -0.07251 to 0.06237).

Since the estrogen metabolic SNPs examined have previously been associated with breast cancer risk, we estimated the correlation between regression coefficients of SNP effects on mammographic density and the odds ratios of SNP effects on breast cancer risk, in order to assess whether the SNPs act through mammographic density as an intermediate phenotype for breast cancer. No significant relationship was found between SNP effects on breast cancer risk and percentage density (Spearman's correlation rho = 0.0411, *P* = 0.5268). Pathway-based multi-SNP association analyses

**Table 2 Significant SNPs in the estrogen metabolic pathway, corresponding regression coefficients and *P* values**

| SNP | Gene | Minor allele | MAF | *n* | Coefficient | SE | *P* value |
|---|---|---|---|---|---|---|---|
| rs11638442 | CYP11A1 | C | 0.35 | 1,677 | 0.0557 | 0.0212 | 0.0088 |
| rs16968478 | CYP11A1 | G | 0.17 | 1,703 | 0.0575 | 0.0263 | 0.0293 |
| rs2279357 | CYP11A1 | A | 0.20 | 1,699 | 0.0511 | 0.0229 | 0.0260 |
| rs2959003 | CYP11A1 | A | 0.28 | 1,669 | 0.0582 | 0.0224 | 0.0094 |
| rs2959008 | CYP11A1 | A | 0.30 | 1,703 | 0.0475 | 0.0221 | 0.0315 |
| rs2066485 | HSD17B3 | G | 0.14 | 1,703 | 0.0668 | 0.0293 | 0.0230 |
| rs7039978 | HSD17B3 | A | 0.50 | 1,694 | −0.0632 | 0.0203 | 0.0019 |
| rs1469908 | NQO1 | C | 0.37 | 1,695 | −0.0472 | 0.0206 | 0.0223 |
| rs17268974 | STS | A | 0.22 | 1,605 | 0.0503 | 0.0238 | 0.0349 |
| rs2270112 | STS | C | 0.34 | 1,686 | −0.0485 | 0.0208 | 0.0197 |
| rs707762 | STS | A | 0.40 | 1,687 | 0.0435 | 0.0205 | 0.0340 |

*P* values from a one-degree-of-freedom likelihood ratio test. MAF, minor allele frequency; SE, standard error.

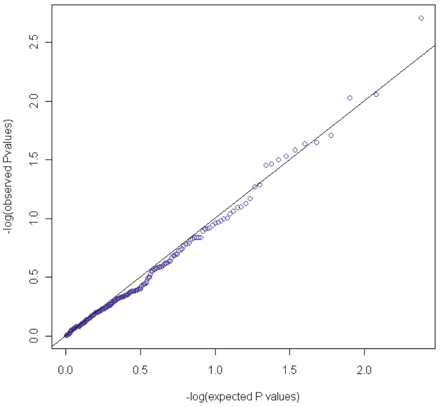


**Figure 2 No association between SNPs and percentage density**. -$\log_{10}$ quantile-quantile *P* value plots from single-SNP trend tests of 239 SNPs in the estrogen metabolism pathway.

**Table 3 Global genetic association tests between SNPs in the estrogen metabolic pathways and mammographic breast density**

| Pathway | Number of SNPs | *P* heterogeneity | *P* trend[a] |
|---|---|---|---|
| Whole pathway | 239 | 0.840 | 0.507 |
| Androgen synthesis | 11 | 0.761 | 0.763 |
| Androgen to estrogen conversion | 120 | 0.587 | 0.715 |
| Estrogen removal | 134 | 0.834 | 0.872 |

[a] *P* values based on 5,000 permutations.

revealed no significant association between percentage density and genetic variations in the overall estrogen metabolic pathway, or any of the related subpathways (Table 3).

## Discussion

Our study suggests there is no appreciable effect between genetic variants involved in estrogen metabolism and mammographic density. Neither the overall estrogen metabolic pathway nor the androgen synthesis, androgen-to-estrogen conversion and estrogen removal subpathways were found to be significantly associated with mammographic density. Single SNP markers with significant associations with mammographic density were not validated in two independent datasets.

In view of estrogen exposure being a major risk factor of postmenopausal breast cancer, and mammographic density being associated with several hormone-related factors such as body mass index (increased local estrogen conversion due to increased fatty tissue), HRT, and menopausal status, the estrogen metabolic pathway has been a candidate pathway for the search of genetic variants related to mammographic density. Most of the variants in the candidate breast cancer genes evaluated in previous studies, however, have been concluded to be only weak predictors of mammographic density [10]. Association findings have been both supported and contradicted [3]. As Boyd and colleagues have discussed [34], it is likely that hormone-related factors are responsible for only a small proportion of the wide variation in mammographic density. In addition, genetic variants involved in the estrogen metabolic pathway are generally investigated based on the premise that mammographic density is an intermediate and heritable risk factor of breast cancer [4]. There is, however, accumulating evidence that mammographic density may predispose to breast cancer risk through components largely independent of estrogen metabolism [35-37].

In our study, no correlation was found between the estimates of SNP effects on breast cancer risk and mammographic density, suggesting that the same SNPs associated with breast cancer risk are not directly correlated with mammographic density. Tamimi and colleagues reported that mammographic density and circulating sex steroid levels were independently associated with breast cancer risk in postmenopausal women [35]. In addition, Kerlikowske and colleagues found no correlation between mammographic density and bone mineral density [36], both of which have been suggested to be cumulative markers of elevated estrogen exposure. Dite and colleagues performed a similar study investigating the overlap between genetic determinants of mammographic density and bone mineral density, and reported a null finding [37]. Another finding in Kerlikowske and colleagues' study was that although mammographic density remained strongly associated with elevated breast cancer risk after adjustment for hormone-related factors, the effects of bone mineral density did not [36], suggesting that estrogen metabolism plays only a small role in the effects of mammographic density on breast cancer risk.

Many studies examining the effects of exogenous estrogen exposure are in agreement with the view that estrogen has limited effects on mammographic density. Very often, the combined estrogen plus progestin regimen was found to affect mammographic density more than the estrogen-only regimen [38-41], suggesting that progestins and not estrogens are responsible for increased mammographic density. Interestingly, mammographic density is also known to have no prognostic bearing on the estrogen receptor status of breast cancer tumors [42-44], thus corroborating an estrogen/estrogen receptor independent link. Another study conducted by Vachon and colleagues found no influence of aromatase inhibitors (drugs that stop the production of estrogen in postmenopausal women) on mammographic density [45], further supporting this line of rationale.

Strengths of the present study include the large sample size and extensive coverage of SNPs in the estrogen metabolic pathway. In a review by Kelemen and colleagues, the authors summarized that previous genetic association studies exploring the relationship between the estrogen metabolic pathway and mammographic density had sample sizes ranging from between 232 and 1,260 women [3]. The number of loci involved in the estrogen metabolic pathway investigated in these studies was also limited to eight or less [3], while we examined 239 tagSNPs from 34 genes involved in the estrogen metabolic pathway. A second strength of the present study is the use of two independent populations for the validation of the associations found.

A limitation of the present work is that it includes different mammogram views across the different studies. The main study on Swedish women utilized the medio-lateral oblique view, while mammograms of the CGEMS and of the Mayo Clinic were taken using the CC view. Several studies, however, have shown correlation of densities from the medio-lateral oblique and CC views [46,47], and have shown that the different views yield similar associations with breast cancer [32]. In addition, the main focus of this study was on genetic determinants of mammographic density in postmenopausal women. Although no strong association was observed between SNPs in the estrogen metabolic pathway examined and mammographic density in postmenopausal women, whether the same lack of association between common genetic variation in the estrogen metabolism pathway and mammographic density is present in premenopausal women remains to be clarified.

## Conclusions

As mammographic density is generally considered an intermediate phenotype for breast cancer, the identification of genes that influence mammographic density would play an important role in risk prediction of breast cancer prior to the start of mammography screenings and shed light on the mechanisms behind breast cancer carcinogenesis. Overall, there is no conclusive evidence that genetic variants in genes involved in the estrogen metabolic pathway are associated with mammographic density in postmenopausal women. This knowledge will be helpful for directing the focus of future studies to alternative pathways that may be responsible for a larger bulk of the genetic component of mammographic density.

**Additional file 1: Tables S1 to S3**. Table S1 presents a list of SNPs in the androgen synthesis subpathway and their corresponding regression coefficients and likelihood ratio test *P* values. Table S2 presents a list of SNPs in the androgen to estrogen conversion subpathway and their corresponding regression coefficients and likelihood ratio test *P* values. Table S3 presents a list of SNPs in the estrogen removal subpathway and their corresponding regression coefficients and likelihood ratio test *P* values.
Click here for file
[ http://www.biomedcentral.com/content/supplementary/bcr2488-S1.DOC ]

**Additional file 2: Table S4**. Table S4 presents genes containing polymorphisms within the estrogen metabolic pathway evaluated in relation to mammographic density.
Click here for file
[ http://www.biomedcentral.com/content/supplementary/bcr2488-S2.DOC ]

**Additional file 3: Table S5**. Table S5 presents validation results of significantly associated SNPs in the Nurses' Health Study (NHS) and the Mayo Clinic Breast Cancer Study (MBCS).
Click here for file
[ http://www.biomedcentral.com/content/supplementary/bcr2488-S3.DOC ]

### Abbreviations

CC: craniocaudal; CGEMS: Cancer Genetic Markers of Susceptibility Project; HRT: hormone replacement therapy; SNP: single nucleotide polymorphism; tagSNP: tagging single nucleotide polymorphism.

### Acknowledgements

The present work was supported by National Institutes of Health (RO1 CA58427), by Märit and Hans Rausing's Initiative Against Breast Cancer, and by the W81XWH-05-1-0314 Innovator Award, US Department of Defense Breast Cancer Research Program, Office of the Congressionally Directed Medical Research Programs. JLi is a recipient of the A*STAR Graduate Scholarship. KH was supported by the Swedish Research Council (523-2006-972). KC was supported by the Swedish Cancer Society (5128-B07-01PAF). Validation work on CGEMS data by the Channing Laboratory was supported by the National Institutes of Health (P01 CA87969, R01 CA065725, R01 CA131332). Validation work at the Mayo Clinic was supported by R01 CA 128931 and R01 CA 122340.

### Author details

[1]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Box 281, 171 77 Stockholm, Sweden. [2]Human Genetics, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore. [3]Channing Laboratory, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115, USA. [4]Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. [5]Program in Genetic and Molecular Epidemiology, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. [6]Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA. [7]Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA. [8]Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, 75 Mikras Asias Str, Goudi, Athens 115 27, Greece.

### Authors' contributions

JLi participated in the study design, carried out the analyses and drafted the manuscript. LE digitized and obtained readings for the mammograms. RMT, SL, DJH, CMV, FJC and CGS contributed to the validation of this study. PL coordinated the Innovator project which contributed data on birthweight and mammographic density. JLi, KH, KC, JLiu and PH conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

Received: 29 December 2009 Revised: 1 March 2010
Accepted: 9 March 2010 Published: 9 March 2010

### References

1. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ: **Mammographic densities and breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 1998, **7**:1133-1144.
2. Vachon CM, van Gils CH, Sellers TA, Ghosh K, Pruthi S, Brandt KR, Pankratz VS: **Mammographic density, breast cancer risk and risk prediction.** *Breast Cancer Res* 2007, **9**:217.
3. Kelemen LE, Sellers TA, Vachon CM: **Can genes for mammographic density inform cancer aetiology?.** *Nat Rev Cancer* 2008, **8**:812-823.
4. Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR, Giles GG, Tritchler D, Chiarelli A, Yaffe MJ, Hopper JL: **Heritability of mammographic density, a risk factor for breast cancer.** *N Engl J Med* 2002, **347**:886-894.
5. Stone J, Gurrin LC, Byrnes GB, Schroen CJ, Treloar SA, Padilla EJ, Dite GS, Southey MC, Hayes VM, Hopper JL: **Mammographic density and candidate gene variants: a twins and sisters study.** *Cancer Epidemiol Biomarkers Prev* 2007, **16**:1479-1484.
6. Boyd NF, Martin LJ, Rommens JM, Paterson AD, Minkin S, Yaffe MJ, Stone J, Hopper JL: **Mammographic density: a heritable risk factor for breast cancer.** *Methods Mol Biol* 2009, **472**:343-360.
7. Boyd NF, Martin LJ, Yaffe MJ, Minkin S: **Mammographic density: a hormonally responsive risk factor for breast cancer.** *J Br Menopause Soc* 2006, **12**:186-193.
8. Duffy SW, Jakes RW, Ng FC, Gao F: **Interaction of dense breast patterns with other breast cancer risk factors in a case-control study.** *Br J Cancer* 2004, **91**:233-236.
9. Martin LJ, Boyd NF: **Mammographic density. Potential mechanisms of breast cancer risk associated with mammographic density: hypotheses based on epidemiological evidence.** *Breast Cancer Res* 2008, **10**:201.
10. Haiman CA, Hankinson SE, De Vivo I, Guillemette C, Ishibe N, Hunter DJ, Byrne C: **Polymorphisms in steroid hormone pathway genes and mammographic density.** *Breast Cancer Res Treat* 2003, **77**:27-36.
11. Maskarinec G, Lurie G, Williams AE, Le Marchand L: **An investigation of mammographic density and gene variants in healthy women.** *Int J Cancer* 2004, **112**:683-688.
12. Lord SJ, Mack WJ, Berg Van Den D, Pike MC, Ingles SA, Haiman CA, Wang W, Parisky YR, Hodis HN, Ursin G: **Polymorphisms in genes involved in estrogen and progesterone metabolism and mammographic density changes in women randomized to postmenopausal hormone therapy: results from a pilot study.** *Breast Cancer Res* 2005, **7**:R336-R344.
13. Chambo D, Kemp C, Costa AM, Souza NC, Guerreiro da Silva ID: **Polymorphism in CYP17, GSTM1 and the progesterone receptor genes and its relationship with mammographic density.** *Braz J Med Biol Res* 2009, **42**:323-329.
14. Haiman CA, Bernstein L, Berg D, Ingles SA, Salane M, Ursin G: **Genetic determinants of mammographic density.** *Breast Cancer Res* 2002, **4**:R5.
15. van Duijnhoven FJ, Bezemer ID, Peeters PH, Roest M, Uitterlinden AG, Grobbee DE, van Gils CH: **Polymorphisms in the estrogen receptor alpha gene and mammographic density.** *Cancer Epidemiol Biomarkers Prev* 2005, **14**:2655-2660.
16. Wedren S, Lovmar L, Humphreys K, Magnusson C, Melhus H, Syvanen AC, Kindmark A, Landegren U, Fermer ML, Stiger F, Persson I, Baron J, Weiderpass E: **Oestrogen receptor alpha gene haplotype and postmenopausal breast cancer risk: a case control study.** *Breast Cancer Res* 2004, **6**:R437-R449.
17. Tamimi RM, Eriksson L, Lagiou P, Czene K, Ekbom A, Hsieh CC, Adami HO, Trichopoulos D, Hall P: **Birth weight and mammographic density among postmenopausal women in Sweden.** *Int J Cancer* **126**:985-991.
18. Byng JW, Boyd NF, Little L, Lockwood G, Fishell E, Jong RA, Yaffe MJ: **Symmetry of projection in the quantitative analysis of mammographic images.** *Eur J Cancer Prev* 1996, **5**:319-327.
19. Qin ZS, Niu T, Liu JS: **Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms.** *Am J Hum Genet* 2002, **71**:1242-1247.
20. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC: **Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study.** *Hum Hered* 2003, **55**:27-36.
21. Tyrer J, Pharoah PD, Easton DF: **The admixture maximum likelihood test: a novel experiment-wise test of association between disease and multiple SNPs.** *Genet Epidemiol* 2006, **30**:636-643.
22. R Development Core Team: *R: A Language and Environment for Statistical Computing* Austria: R Foundation for Statistical Computing 2008.
23. Everett BM, Kurth T, Buring JE, Ridker PM: **The relative strength of C-reactive protein and lipid levels as determinants of ischemic stroke compared with coronary heart disease in women.** *J Am Coll Cardiol* 2006, **48**:2235-2242.
24. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.** *Nat Genet* 2007, **39**:870-874.
25. Colditz GA, Hankinson SE: **The Nurses' Health Study: lifestyle and health among women.** *Nat Rev Cancer* 2005, **5**:388-396.
26. Hankinson SE, Willett WC, Manson JE, Colditz GA, Hunter DJ, Spiegelman D, Barbieri RL, Speizer FE: **Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women.** *J Natl Cancer Inst* 1998, **90**:1292-1299.
27. Tamimi RM, Cox DG, Kraft P, Pollak MN, Haiman CA, Cheng I, Freedman ML, Hankinson SE, Hunter DJ, Colditz GA: **Common genetic variation in IGF1, IGFBP-1, and IGFBP-3 in relation to mammographic density: a cross-sectional study.** *Breast Cancer Res* 2007, **9**:R18.
28. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation.** *Annu Rev Genomics Hum Genet* 2009, **10**:387-406.
29. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 2007, **23**:1294-1296.
30. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci.** *Nature* 2007, **447**:1087-1093.
31. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S, Lissowska J, Brinton L, Peplonska B, Southey MC, Hopper JL, McCredie MR, Giles GG, Fletcher O, Johnson N, dos Santos Silva I, Gibson L, Bojesen SE, Nordestgaard BG, Axelsson CK, Torres D, Hamann U, Justenhoven C, Brauch H, Chang-Claude J, Kropp S, *et al*: **A common coding variant in CASP8 is associated with breast cancer risk.** *Nat Genet* 2007, **39**:352-358.
32. Vachon CM, Brandt KR, Ghosh K, Scott CG, Maloney SD, Carston MJ, Pankratz VS, Sellers TA: **Mammographic breast density as a general marker of breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 2007, **16**:43-49.
33. Boyd NF, Stone J, Martin LJ, Jong R, Fishell E, Yaffe M, Hammond G, Minkin S: **The association of breast mitogens with mammographic densities.** *Br J Cancer* 2002, **87**:876-882.
34. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, Paterson AD: **Mammographic breast density as an intermediate phenotype for breast cancer.** *Lancet Oncol* 2005, **6**:798-808.
35. Tamimi RM, Byrne C, Colditz GA, Hankinson SE: **Endogenous hormone levels, mammographic density, and subsequent risk of breast cancer in postmenopausal women.** *J Natl Cancer Inst* 2007, **99**:1178-1187.
36. Kerlikowske K, Shepherd J, Creasman J, Tice JA, Ziv E, Cummings SR: **Are breast density and bone mineral density independent risk factors for breast cancer?** *J Natl Cancer Inst* 2005, **97**:368-374.
37. Dite GS, Wark JD, Giles GG, English DR, McCredie MR, Hopper JL: **Is there overlap between the genetic determinants of mammographic density and bone mineral density?** *Cancer Epidemiol Biomarkers Prev* 2005, **14**:2266-2268.
38. Kaewrudee S, Anuwutnavin S, Kanpittaya J, Soontrapa S, Sakondhavat C: **Effect of estrogen-progestin and estrogen on mammographic density.** *J Reprod Med* 2007, **52**:513-520.

39. McTiernan A, Martin CF, Peck JD, Aragaki AK, Chlebowski RT, Pisano ED, Wang CY, Brunner RL, Johnson KC, Manson JE, Lewis CE, Kotchen JM, Hulka BS: **Estrogen-plus-progestin use and mammographic density in postmenopausal women: Women's Health Initiative randomized trial.** *J Natl Cancer Inst* 2005, **97**:1366-1376.
40. Topal NB, Ayhan S, Topal U, Bilgin T: **Effects of hormone replacement therapy regimens on mammographic breast density: the role of progestins.** *J Obstet Gynaecol Res* 2006, **32**:305-308.
41. McTiernan A, Chlebowski RT, Martin C, Peck JD, Aragaki A, Pisano ED, Wang CY, Johnson KC, Manson JE, Wallace RB, Vitolins MZ, Heiss G: **Conjugated equine estrogen influence on mammographic density in postmenopausal women in a substudy of the women's health initiative randomized trial.** *J Clin Oncol* 2009, **27**:6135-6143.
42. Ghosh K, Brandt KR, Sellers TA, Reynolds C, Scott CG, Maloney SD, Carston MJ, Pankratz VS, Vachon CM: **Association of mammographic density with the pathology of subsequent breast cancer among postmenopausal women.** *Cancer Epidemiol Biomarkers Prev* 2008, **17**:872-879.
43. Chen JH, Hsu FT, Shih HN, Hsu CC, Chang D, Nie K, Nalcioglu O, Su MY: **Does breast density show difference in patients with estrogen receptor-positive and estrogen receptor-negative breast cancer measured on MRI?** *Ann Oncol* 2009, **20**:1447-1449.
44. Ziv E, Tice J, Smith-Bindman R, Shepherd J, Cummings S, Kerlikowske K: **Mammographic density and estrogen receptor status of breast cancer.** *Cancer Epidemiol Biomarkers Prev* 2004, **13**:2090-2095.
45. Vachon CM, Ingle JN, Suman VJ, Scott CG, Gottardt H, Olson JE, Goss PE: **Pilot study of the impact of letrozole vs. placebo on breast density in women completing 5 years of tamoxifen.** *Breast* 2007, **16**:204-210.
46. 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.
47. Kim SJ, Moon WK, Cho N, Cha JH, Kim SM, Im JG: **Computer-aided detection in digital mammography: comparison of craniocaudal, mediolateral oblique, and mediolateral views.** *Radiology* 2006, **241**:695-701.

doi:10.1186/bcr2488
**Cite this article as:** Li *et al.*: Genetic variation in the estrogen metabolic pathway and mammographic density as an intermediate phenotype of breast cancer. *Breast Cancer Research* 2010 **12**:R19.

**Table S1. List of SNPs in the androgen synthesis sub-pathway and their corresponding regression coefficients and likelihood ratio test pvalues.**

| Chr | Position | SNP | Gene | N | Coefficient | SE | P |
|---|---|---|---|---|---|---|---|
| 10 | 104581383 | rs17115100 | CYP17A1 | 1665 | 0.0120 | 0.0325 | 0.7130 |
| 10 | 104584497 | rs1004467 | CYP17A1 | 1696 | -0.0148 | 0.0327 | 0.6503 |
| 10 | 104585709 | rs3781286 | CYP17A1 | 1684 | -0.0222 | 0.0206 | 0.2805 |
| 10 | 104587470 | rs2486758 | CYP17A1 | 1675 | 0.0111 | 0.0253 | 0.6601 |
| 10 | 104595318 | rs7089422 | CYP17A1 | 1692 | 0.0196 | 0.0262 | 0.4553 |
| 15 | 72403630 | rs2959008 | CYP11A1 | 1703 | 0.0475 | 0.0221 | 0.0315 |
| 15 | 72415944 | rs2959003 | CYP11A1 | 1669 | 0.0582 | 0.0224 | 0.0094 |
| 15 | 72417676 | rs2279357 | CYP11A1 | 1699 | 0.0511 | 0.0229 | 0.0260 |
| 15 | 72421952 | rs11638442 | CYP11A1 | 1677 | 0.0557 | 0.0212 | 0.0088 |
| 15 | 72449864 | rs16968478 | CYP11A1 | 1703 | 0.0575 | 0.0263 | 0.0293 |
| 15 | 72451904 | rs8039957 | CYP11A1 | 1705 | 0.0356 | 0.0303 | 0.2401 |

Chr: chromosome; SNP: single nucleotide polymorphism rsid; N: number of subjects; SE: standard error; P: P-value for 1 d.f. trend test

**Table S2. List of SNPs in the androgen to estrogen conversion sub-pathway and their corresponding regression coefficients and likelihood ratio test pvalues.**

| Chr | Position | SNP | Gene | N | Coefficient | SE | P |
|---|---|---|---|---|---|---|---|
| 1 | 119752771 | rs6428822 | HSD3B1 | 1585 | 0.0275 | 0.0209 | 0.1891 |
| 1 | 119757996 | rs4659175 | HSD3B1 | 1676 | 0.0154 | 0.0217 | 0.4775 |
| 1 | 119783549 | rs1341013 | HSD3B1 | 1692 | 0.0331 | 0.0206 | 0.1077 |
| 1 | 119800735 | rs6672903 | HSD3B1 | 1656 | -0.0163 | 0.0205 | 0.4248 |
| 1 | 119810999 | rs2298029 | HSD3B1 | 1698 | 0.0262 | 0.0219 | 0.2320 |
| 1 | 119826497 | rs911245 | HSD3B1 | 1646 | 0.0244 | 0.0218 | 0.2627 |
| 1 | 119861023 | rs10923844 | HSD3B1 | 1658 | 0.0180 | 0.0224 | 0.4221 |
| 1 | 207918699 | rs11576775 | HSD11B1 | 1644 | 0.0269 | 0.0259 | 0.2980 |
| 1 | 207925076 | rs846908 | HSD11B1 | 1694 | -0.0003 | 0.0608 | 0.9955 |
| 1 | 207933739 | rs10082248 | HSD11B1 | 1699 | -0.0040 | 0.0462 | 0.9303 |
| 1 | 207937539 | rs4844880 | HSD11B1 | 1689 | -0.0160 | 0.0254 | 0.5288 |
| 1 | 207948638 | rs2282738 | HSD11B1 | 1687 | -0.0107 | 0.0233 | 0.6442 |
| 1 | 207951994 | rs968033 | HSD11B1 | 1698 | -0.0049 | 0.0541 | 0.9283 |
| 1 | 207954341 | rs846906 | HSD11B1 | 1692 | 0.0053 | 0.0278 | 0.8487 |
| 1 | 207989777 | rs6702301 | HSD11B1 | 1687 | -0.0116 | 0.0230 | 0.6155 |
| 1 | 207996076 | rs2272866 | HSD11B1 | 1694 | 0.0521 | 0.0815 | 0.5227 |
| 2 | 31596366 | rs2208158 | SRD5A2 | 1696 | 0.0239 | 0.0220 | 0.2770 |
| 2 | 31602532 | rs3731586 | SRD5A2 | 1667 | 0.0498 | 0.0358 | 0.1635 |
| 2 | 31617062 | rs12470143 | SRD5A2 | 1627 | -0.0231 | 0.0206 | 0.2614 |
| 2 | 31620635 | rs4952197 | SRD5A2 | 1642 | 0.0322 | 0.0223 | 0.1483 |
| 2 | 31635784 | rs2268796 | SRD5A2 | 1636 | 0.0365 | 0.0208 | 0.0794 |
| 2 | 31640141 | rs2300697 | SRD5A2 | 1653 | 0.0090 | 0.0207 | 0.6627 |
| 2 | 31651315 | rs6749019 | SRD5A2 | 1616 | 0.0079 | 0.0207 | 0.7047 |
| 2 | 234175427 | rs2741019 | UGT1A1.9 | 1699 | 0.0169 | 0.0224 | 0.4521 |
| 2 | 234201376 | rs1377460 | UGT1A1.9 | 1709 | -0.0370 | 0.0254 | 0.1459 |
| 2 | 234251553 | rs7587916 | UGT1A1.9 | 1670 | -0.0142 | 0.0207 | 0.4919 |
| 2 | 234282371 | rs4663327 | UGT1A1.9 | 1691 | -0.0161 | 0.0333 | 0.6295 |
| 2 | 234295182 | rs7597496 | UGT1A1.9 | 1573 | -0.0043 | 0.0205 | 0.8321 |
| 2 | 234330521 | rs10929302 | UGT1A1.9 | 1621 | 0.0053 | 0.0226 | 0.8128 |
| 2 | 234337378 | rs6742078 | UGT1A1.9 | 1695 | 0.0047 | 0.0215 | 0.8262 |
| 2 | 234346283 | rs1042640 | UGT1A1.9 | 1709 | -0.0186 | 0.0254 | 0.4634 |
| 2 | 234348089 | rs11563250 | UGT1A1.9 | 1695 | 0.0079 | 0.0289 | 0.7831 |
| 2 | 234348502 | rs6719561 | UGT1A1.9 | 1681 | 0.0042 | 0.0215 | 0.8461 |
| 2 | 234367644 | rs10169532 | UGT1A1.9 | 1593 | -0.0147 | 0.0205 | 0.4737 |
| 2 | 234371560 | hCV256966 | UGT1A1.9 | 1677 | 0.0151 | 0.0228 | 0.5096 |
| 4 | 69904593 | rs11932983 | UGT2B11 | 1677 | 0.0203 | 0.0287 | 0.4805 |
| 4 | 69910216 | rs2331627 | UGT2B11 | 1666 | -0.0242 | 0.0262 | 0.3558 |
| 4 | 69966297 | rs10030066 | UGT2B11 | 1639 | 0.0150 | 0.0243 | 0.5363 |
| 4 | 69975780 | rs7677996 | UGT2B11 | 1602 | -0.0417 | 0.0216 | 0.0537 |
| 4 | 70024587 | rs4371687 | UGT2B11 | 1686 | -0.0110 | 0.0196 | 0.5749 |
| 4 | 70041206 | rs6837285 | UGT2B11 | 1674 | 0.0144 | 0.0196 | 0.4620 |
| 4 | 70075861 | rs6600903 | UGT2B11 | 1678 | -0.0329 | 0.0203 | 0.1053 |
| 4 | 70370761 | rs2736520 | UGT2B4 | 1660 | -0.0170 | 0.0287 | 0.5542 |
| 4 | 70370923 | rs903445 | UGT2B4 | 1663 | 0.0085 | 0.0202 | 0.6746 |
| 4 | 70375511 | rs1494798 | UGT2B4 | 1660 | 0.0173 | 0.0212 | 0.4151 |
| 4 | 70379230 | rs1080755 | UGT2B4 | 1601 | -0.0264 | 0.0239 | 0.2711 |
| 4 | 70389067 | rs2013573 | UGT2B4 | 1696 | -0.0285 | 0.0247 | 0.2483 |
| 4 | 70394283 | rs7441743 | UGT2B4 | 1528 | 0.0034 | 0.0211 | 0.8718 |
| 4 | 70397951 | rs6600771 | UGT2B4 | 1609 | 0.0456 | 0.0284 | 0.1092 |
| 5 | 6690380 | rs531241 | SRD5A1 | 1701 | 0.0202 | 0.0202 | 0.3174 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | 6708247 | rs568509 | SRD5A1 | 1614 | 0.0047 | 0.0302 | 0.8770 |
| 5 | 6718364 | rs4702381 | SRD5A1 | 1689 | 0.0061 | 0.0239 | 0.7973 |
| 5 | 6723649 | rs16877779 | SRD5A1 | 1674 | -0.0493 | 0.0338 | 0.1454 |
| 5 | 6734187 | rs768437 | SRD5A1 | 1694 | -0.0161 | 0.0244 | 0.5107 |
| 8 | 143947798 | rs4464947 | CYP11B1 | 1696 | 0.0070 | 0.0353 | 0.8426 |
| 8 | 143952659 | rs5297 | CYP11B1 | 1701 | -0.0074 | 0.0353 | 0.8347 |
| 8 | 143989866 | rs3802230 | CYP11B2 | 1698 | 0.0163 | 0.0198 | 0.4102 |
| 8 | 143990317 | rs3097 | CYP11B2 | 1686 | 0.0064 | 0.0218 | 0.7699 |
| 8 | 143992745 | rs4543 | CYP11B2 | 1713 | 0.0058 | 0.0355 | 0.8706 |
| 8 | 143996602 | rs1799998 | CYP11B2 | 1658 | -0.0150 | 0.0196 | 0.4450 |
| 10 | 5224291 | rs1334466 | AKR1C4 | 1691 | -0.0080 | 0.0212 | 0.7060 |
| 10 | 5228196 | rs4880716 | AKR1C4 | 1684 | 0.0253 | 0.0224 | 0.2577 |
| 10 | 5233017 | rs7085249 | AKR1C4 | 1680 | 0.0203 | 0.0225 | 0.3653 |
| 10 | 5234441 | rs2151896 | AKR1C4 | 1690 | 0.0115 | 0.0199 | 0.5642 |
| 10 | 5237376 | rs3750572 | AKR1C4 | 1703 | 0.0224 | 0.0252 | 0.3752 |
| 10 | 5239111 | rs4881412 | AKR1C4 | 1698 | 0.0137 | 0.0362 | 0.7040 |
| 10 | 5240453 | rs1931679 | AKR1C4 | 1697 | 0.0384 | 0.0311 | 0.2160 |
| 10 | 5244821 | rs1831977 | AKR1C4 | 1623 | 0.0143 | 0.0264 | 0.5888 |
| 10 | 5246185 | rs12762017 | AKR1C4 | 1616 | -0.0184 | 0.0293 | 0.5313 |
| 10 | 5246497 | rs17134588 | AKR1C4 | 1646 | 0.0205 | 0.0260 | 0.4305 |
| 10 | 5248069 | rs10458795 | AKR1C4 | 1702 | 0.0181 | 0.0524 | 0.7303 |
| 15 | 49279146 | rs9972359 | CYP19A1 | 1684 | 0.0119 | 0.0203 | 0.5561 |
| 15 | 49283122 | rs934632 | CYP19A1 | 1690 | 0.0137 | 0.0253 | 0.5869 |
| 15 | 49286837 | rs7167936 | CYP19A1 | 1693 | 0.0173 | 0.0204 | 0.3957 |
| 15 | 49290136 | rs4646 | CYP19A1 | 1694 | 0.0034 | 0.0223 | 0.8775 |
| 15 | 49301213 | rs959564 | CYP19A1 | 1694 | -0.0327 | 0.0393 | 0.4059 |
| 15 | 49304392 | rs12595627 | CYP19A1 | 1657 | 0.0153 | 0.0217 | 0.4797 |
| 15 | 49324419 | hCV8234885 | CYP19A1 | 1628 | 0.0165 | 0.0208 | 0.4275 |
| 15 | 49344549 | rs12050767 | CYP19A1 | 1643 | -0.0228 | 0.0205 | 0.2667 |
| 15 | 49379835 | rs17523880 | CYP19A1 | 1697 | -0.0247 | 0.0307 | 0.4217 |
| 15 | 49382264 | hCV3060064 | CYP19A1 | 1680 | 0.0230 | 0.0205 | 0.2611 |
| 15 | 49383831 | rs8031463 | CYP19A1 | 1702 | 0.0249 | 0.0466 | 0.5927 |
| 15 | 49393870 | rs3751592 | CYP19A1 | 1670 | 0.0058 | 0.0218 | 0.7912 |
| 15 | 49397006 | rs2470150 | CYP19A1 | 1710 | -0.0271 | 0.0403 | 0.5006 |
| 15 | 49401198 | rs1902585 | CYP19A1 | 1707 | -0.0044 | 0.0205 | 0.8317 |
| 16 | 65992902 | rs11642680 | HSD11B2 | 1712 | -0.0538 | 0.0635 | 0.3968 |
| 16 | 65994359 | rs2059237 | HSD11B2 | 1702 | 0.0255 | 0.0529 | 0.6297 |
| 16 | 66007332 | rs7206718 | HSD11B2 | 1680 | 0.0237 | 0.0202 | 0.2409 |
| 16 | 66011135 | rs8047159 | HSD11B2 | 1701 | 0.0367 | 0.0323 | 0.2562 |
| 16 | 66029427 | rs4360931 | HSD11B2 | 1700 | 0.0269 | 0.0383 | 0.4831 |
| 16 | 66049072 | rs749242 | HSD11B2 | 1697 | 0.0279 | 0.0383 | 0.4660 |
| 19 | 53037973 | rs7248427 | SULT2A1 | 1675 | 0.0073 | 0.0205 | 0.7207 |
| 19 | 53040066 | rs17239147 | SULT2A1 | 1686 | -0.0033 | 0.0289 | 0.9084 |
| 19 | 53048964 | rs4483956 | SULT2A1 | 1688 | 0.0282 | 0.0207 | 0.1720 |
| 19 | 53063945 | rs188440 | SULT2A1 | 1706 | 0.0047 | 0.0228 | 0.8359 |
| 19 | 53067510 | rs296364 | SULT2A1 | 1678 | -0.0034 | 0.0196 | 0.8627 |
| 19 | 53074328 | rs11083905 | SULT2A1 | 1706 | -0.0181 | 0.0336 | 0.5897 |
| 19 | 53083388 | rs7508610 | SULT2A1 | 1642 | 0.0005 | 0.0210 | 0.9797 |
| 19 | 53090448 | rs2972612 | SULT2A1 | 1648 | 0.0002 | 0.0227 | 0.9942 |
| 19 | 53745118 | rs279451 | SULT2B1 | 1659 | -0.0005 | 0.0287 | 0.9875 |
| 19 | 53747608 | rs279447 | SULT2B1 | 1715 | 0.0142 | 0.0482 | 0.7680 |
| 19 | 53753536 | rs3848542 | SULT2B1 | 1694 | -0.0318 | 0.0227 | 0.1613 |
| 19 | 53756771 | rs12611137 | SULT2B1 | 1700 | 0.0005 | 0.0259 | 0.9858 |
| 19 | 53762686 | rs2665605 | SULT2B1 | 1703 | -0.0128 | 0.0300 | 0.6703 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 19 | 53766894 | rs2665577 | SULT2B1 | 1701 | 0.0057 | 0.0212 | 0.7865 |
| 19 | 53775305 | rs6509396 | SULT2B1 | 1700 | -0.0125 | 0.0210 | 0.5525 |
| 19 | 53784242 | rs10426628 | SULT2B1 | 1679 | 0.0335 | 0.0234 | 0.1522 |
| 19 | 53791303 | rs2665587 | SULT2B1 | 1706 | 0.0085 | 0.0282 | 0.7629 |
| 19 | 53791767 | rs3815691 | SULT2B1 | 1706 | 0.0382 | 0.0298 | 0.2003 |
| 19 | 53794211 | rs1132054 | SULT2B1 | 1679 | 0.0071 | 0.0203 | 0.7282 |
| 19 | 53812246 | rs369880 | SULT2B1 | 1691 | -0.0108 | 0.0261 | 0.6799 |
| 23 | 7125093 | rs707762 | STS | 1687 | 0.0435 | 0.0205 | 0.0340 |
| 23 | 7180925 | rs2270112 | STS | 1686 | -0.0485 | 0.0208 | 0.0197 |
| 23 | 7184199 | rs12861247 | STS | 1701 | 0.0081 | 0.0328 | 0.8050 |
| 23 | 7194615 | rs5934850 | STS | 1617 | 0.0373 | 0.0209 | 0.0744 |
| 23 | 7224805 | rs5934914 | STS | 1587 | -0.0008 | 0.0231 | 0.9722 |
| 23 | 7246970 | rs17268974 | STS | 1605 | 0.0503 | 0.0238 | 0.0349 |
| 23 | 7253304 | rs4403552 | STS | 1694 | 0.0072 | 0.0245 | 0.7683 |
| 23 | 7264481 | rs17268988 | STS | 1687 | -0.0270 | 0.0228 | 0.2362 |
| 23 | 7280996 | rs1131289 | STS | 1691 | 0.0079 | 0.0220 | 0.7195 |

Chr: chromosome; SNP: single nucleotide polymorphism rsid; N: number of subjects; SE: standard error; P: P-value for 1 d.f. trend test

**Table S3. List of SNPs in estrogen removal sub-pathway and their corresponding regression coefficients and likelihood ratio test pvalues.**

| Chr | Position | SNP | Gene | N | Coefficient | SE | P |
|---|---|---|---|---|---|---|---|
| 1 | 159531389 | hCV2765051 | HSD17B7 | 1662 | -0.0150 | 0.0302 | 0.6188 |
| 1 | 161014649 | rs1780007 | HSD17B7 | 1700 | -0.0230 | 0.0252 | 0.3602 |
| 1 | 161043150 | rs1039874 | HSD17B7 | 1686 | -0.0205 | 0.0485 | 0.6719 |
| 1 | 161061083 | rs1704767 | HSD17B7 | 1676 | 0.0097 | 0.0203 | 0.6338 |
| 1 | 161061423 | rs1006390 | HSD17B7 | 1670 | -0.0115 | 0.0237 | 0.6273 |
| 2 | 38136239 | rs163076 | CYP1B1 | 1672 | -0.0295 | 0.0212 | 0.1634 |
| 2 | 38145208 | rs2256327 | CYP1B1 | 1659 | 0.0250 | 0.0250 | 0.3166 |
| 2 | 38146266 | rs163086 | CYP1B1 | 1677 | 0.0125 | 0.0248 | 0.6150 |
| 2 | 38151707 | rs1056836 | CYP1B1 | 1682 | 0.0047 | 0.0200 | 0.8154 |
| 2 | 38156298 | rs2551188 | CYP1B1 | 1693 | -0.0046 | 0.0218 | 0.8337 |
| 2 | 234175427 | rs2741019 | UGT1A1.9 | 1699 | 0.0169 | 0.0224 | 0.4521 |
| 2 | 234201376 | rs1377460 | UGT1A1.9 | 1709 | -0.0370 | 0.0254 | 0.1459 |
| 2 | 234251553 | rs7587916 | UGT1A1.9 | 1670 | -0.0142 | 0.0207 | 0.4919 |
| 2 | 234282371 | rs4663327 | UGT1A1.9 | 1691 | -0.0161 | 0.0333 | 0.6295 |
| 2 | 234295182 | rs7597496 | UGT1A1.9 | 1573 | -0.0043 | 0.0205 | 0.8321 |
| 2 | 234330521 | rs10929302 | UGT1A1.9 | 1621 | 0.0053 | 0.0226 | 0.8128 |
| 2 | 234337378 | rs6742078 | UGT1A1.9 | 1695 | 0.0047 | 0.0215 | 0.8262 |
| 2 | 234346283 | rs1042640 | UGT1A1.9 | 1709 | -0.0186 | 0.0254 | 0.4634 |
| 2 | 234348089 | rs11563250 | UGT1A1.9 | 1695 | 0.0079 | 0.0289 | 0.7831 |
| 2 | 234348502 | rs6719561 | UGT1A1.9 | 1681 | 0.0042 | 0.0215 | 0.8461 |
| 2 | 234367644 | rs10169532 | UGT1A1.9 | 1593 | -0.0147 | 0.0205 | 0.4737 |
| 2 | 234371560 | hCV256966 | UGT1A1.9 | 1677 | 0.0151 | 0.0228 | 0.5096 |
| 4 | 69904593 | rs11932983 | UGT2B11 | 1677 | 0.0203 | 0.0287 | 0.4805 |
| 4 | 69910216 | rs2331627 | UGT2B11 | 1666 | -0.0242 | 0.0262 | 0.3558 |
| 4 | 69966297 | rs10030066 | UGT2B11 | 1639 | 0.0150 | 0.0243 | 0.5363 |
| 4 | 69975780 | rs7677996 | UGT2B11 | 1602 | -0.0417 | 0.0216 | 0.0537 |
| 4 | 70024587 | rs4371687 | UGT2B11 | 1686 | -0.0110 | 0.0196 | 0.5749 |
| 4 | 70041206 | rs6837285 | UGT2B11 | 1674 | 0.0144 | 0.0196 | 0.4620 |
| 4 | 70075861 | rs6600903 | UGT2B11 | 1678 | -0.0329 | 0.0203 | 0.1053 |
| 4 | 70370761 | rs2736520 | UGT2B4 | 1660 | -0.0170 | 0.0287 | 0.5542 |
| 4 | 70370923 | rs903445 | UGT2B4 | 1663 | 0.0085 | 0.0202 | 0.6746 |
| 4 | 70375511 | rs1494798 | UGT2B4 | 1660 | 0.0173 | 0.0212 | 0.4151 |
| 4 | 70379230 | rs1080755 | UGT2B4 | 1601 | -0.0264 | 0.0239 | 0.2711 |
| 4 | 70389067 | rs2013573 | UGT2B4 | 1696 | -0.0285 | 0.0247 | 0.2483 |
| 4 | 70394283 | rs7441743 | UGT2B4 | 1528 | 0.0034 | 0.0211 | 0.8718 |
| 4 | 70397951 | rs6600771 | UGT2B4 | 1609 | 0.0456 | 0.0284 | 0.1092 |
| 4 | 70736595 | rs1529039 | STE..SULT1E1. | 1685 | -0.0463 | 0.0297 | 0.1197 |
| 4 | 70740955 | rs1220725 | STE..SULT1E1. | 1522 | 0.0032 | 0.0321 | 0.9212 |
| 4 | 70743796 | rs3775779 | STE..SULT1E1. | 1659 | 0.0274 | 0.0216 | 0.2055 |
| 4 | 70752594 | rs4149534 | STE..SULT1E1. | 1690 | -0.0096 | 0.0227 | 0.6732 |
| 4 | 70753566 | rs1220716 | STE..SULT1E1. | 1691 | -0.0843 | 0.0483 | 0.0812 |
| 4 | 70760988 | rs4149525 | STE..SULT1E1. | 1672 | 0.0165 | 0.0278 | 0.5513 |
| 4 | 70774109 | rs1154741 | STE..SULT1E1. | 1700 | -0.0059 | 0.0214 | 0.7830 |
| 5 | 118797197 | rs154632 | HSD17B4 | 1668 | 0.0107 | 0.0220 | 0.6273 |
| 5 | 118799322 | rs13154090 | HSD17B4 | 1703 | -0.0328 | 0.0489 | 0.5023 |
| 5 | 118816919 | rs10478424 | HSD17B4 | 1690 | -0.0072 | 0.0232 | 0.7550 |
| 5 | 118820620 | rs11749784 | HSD17B4 | 1694 | -0.0236 | 0.0239 | 0.3247 |
| 5 | 118830120 | rs1283826 | HSD17B4 | 1697 | -0.0409 | 0.0382 | 0.2836 |
| 5 | 118835035 | rs439954 | HSD17B4 | 1602 | 0.0468 | 0.0303 | 0.1224 |
| 5 | 118860864 | rs3756513 | HSD17B4 | 1688 | -0.0055 | 0.0315 | 0.8607 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 5 | 118904980 | rs17388769 | HSD17B4 | 1695 | -0.0169 | 0.0303 | 0.5782 |
| 6 | 33266876 | rs2269346 | HSD17B8 | 1669 | 0.0160 | 0.0471 | 0.7343 |
| 6 | 33270060 | rs2072915 | HSD17B8 | 1676 | -0.0132 | 0.0222 | 0.5530 |
| 6 | 33277873 | rs1547387 | HSD17B8 | 1708 | 0.0062 | 0.0343 | 0.8567 |
| 6 | 33280910 | rs110662 | HSD17B8 | 1608 | -0.0046 | 0.0219 | 0.8321 |
| 6 | 160018212 | rs4342445 | SOD2 | 1691 | 0.0393 | 0.0238 | 0.0985 |
| 6 | 160020106 | rs2842980 | SOD2 | 1685 | -0.0171 | 0.0251 | 0.4968 |
| 6 | 160023074 | rs5746136 | SOD2 | 1676 | 0.0185 | 0.0220 | 0.3996 |
| 6 | 160027081 | rs1800665 | SOD2 | 1684 | -0.0548 | 0.0956 | 0.5664 |
| 6 | 160030444 | rs2758334 | SOD2 | 1645 | -0.0032 | 0.0200 | 0.8712 |
| 7 | 99013350 | hCV11246907 | CYP3A4_5 | 1705 | 0.0061 | 0.0586 | 0.9174 |
| 7 | 99083016 | rs4646457 | CYP3A4_5 | 1680 | 0.0184 | 0.0386 | 0.6339 |
| 7 | 99104254 | rs4646450 | CYP3A4_5 | 1670 | 0.0143 | 0.0285 | 0.6170 |
| 7 | 99142648 | rs2687078 | CYP3A4_5 | 1691 | 0.0077 | 0.0349 | 0.8254 |
| 7 | 99170019 | rs2687133 | CYP3A4_5 | 1706 | -0.0009 | 0.0379 | 0.9816 |
| 7 | 99186264 | rs6945984 | CYP3A4_5 | 1622 | -0.0127 | 0.0321 | 0.6923 |
| 8 | 18110372 | rs11203942 | NAT1 | 1673 | -0.0115 | 0.0210 | 0.5860 |
| 8 | 18113444 | rs3850751 | NAT1 | 1711 | -0.0152 | 0.0202 | 0.4508 |
| 8 | 18118222 | rs6586714 | NAT1 | 1689 | 0.0508 | 0.0321 | 0.1135 |
| 8 | 18120186 | rs4921880 | NAT1 | 1692 | 0.0012 | 0.0238 | 0.9592 |
| 8 | 18120277 | rs11777998 | NAT1 | 1694 | -0.0150 | 0.0355 | 0.6715 |
| 8 | 18121590 | rs7003890 | NAT1 | 1691 | -0.0100 | 0.0202 | 0.6218 |
| 8 | 18122267 | rs8190837 | NAT1 | 1700 | -0.0269 | 0.0330 | 0.4159 |
| 8 | 18287058 | rs4921906 | NAT2 | 1685 | 0.0268 | 0.0202 | 0.1860 |
| 8 | 18295202 | rs9987109 | NAT2 | 1703 | 0.0232 | 0.0204 | 0.2550 |
| 8 | 18298747 | rs2410556 | NAT2 | 1594 | -0.0012 | 0.0319 | 0.9693 |
| 8 | 18306826 | rs4646257 | NAT2 | 1688 | -0.0243 | 0.0257 | 0.3443 |
| 8 | 18307392 | rs1495748 | NAT2 | 1680 | -0.0077 | 0.0216 | 0.7220 |
| 8 | 18309403 | rs1495738 | NAT2 | 1673 | 0.0355 | 0.0207 | 0.0860 |
| 8 | 18316718 | rs4921914 | NAT2 | 1693 | -0.0011 | 0.0246 | 0.9630 |
| 9 | 98026168 | rs442686 | HSD17B3 | 1662 | -0.0421 | 0.0216 | 0.0513 |
| 9 | 98027222 | rs4306016 | HSD17B3 | 1623 | 0.0067 | 0.0202 | 0.7389 |
| 9 | 98043085 | rs2066485 | HSD17B3 | 1703 | 0.0668 | 0.0293 | 0.0230 |
| 9 | 98058102 | rs8190534 | HSD17B3 | 1685 | 0.0359 | 0.0247 | 0.1460 |
| 9 | 98061403 | rs7039978 | HSD17B3 | 1694 | -0.0632 | 0.0203 | 0.0019 |
| 9 | 98069778 | rs8190530 | HSD17B3 | 1708 | 0.0105 | 0.0203 | 0.6062 |
| 9 | 98091939 | rs7022250 | HSD17B3 | 1696 | -0.0351 | 0.0208 | 0.0915 |
| 9 | 98104670 | rs8190479 | HSD17B3 | 1619 | -0.0058 | 0.0396 | 0.8845 |
| 11 | 67100533 | rs656652 | GSTP1 | 1705 | -0.0108 | 0.0201 | 0.5921 |
| 15 | 72790561 | rs6495121 | CYP1A1.2 | 1680 | 0.0390 | 0.0304 | 0.2008 |
| 15 | 72800040 | rs1799814 | CYP1A1.2 | 1698 | -0.0215 | 0.0563 | 0.7026 |
| 15 | 72806502 | rs2470893 | CYP1A1.2 | 1702 | 0.0161 | 0.0213 | 0.4493 |
| 15 | 72814933 | rs2472297 | CYP1A1.2 | 1665 | 0.0120 | 0.0223 | 0.5910 |
| 15 | 72839115 | rs1350194 | CYP1A1.2 | 1711 | -0.0556 | 0.0625 | 0.3737 |
| 16 | 28507783 | rs17639997 | SULT1A1.2 | 1714 | -0.0244 | 0.0386 | 0.5280 |
| 16 | 28517197 | rs12445705 | SULT1A1.2 | 1562 | -0.0116 | 0.0455 | 0.7994 |
| 16 | 28521466 | rs11074907 | SULT1A1.2 | 1609 | 0.0273 | 0.0204 | 0.1814 |
| 16 | 28523209 | rs11074904 | SULT1A1.2 | 1697 | -0.0230 | 0.0318 | 0.4701 |
| 16 | 28524629 | rs6839 | SULT1A1.2 | 1569 | 0.0168 | 0.0205 | 0.4147 |
| 16 | 28539522 | rs2411453 | SULT1A1.2 | 1618 | 0.0345 | 0.0210 | 0.1004 |
| 16 | 68287295 | rs12595869 | NQO1 | 1698 | -0.0493 | 0.0269 | 0.0672 |
| 16 | 68287927 | rs1437134 | NQO1 | 1637 | 0.0184 | 0.0201 | 0.3614 |
| 16 | 68288056 | rs3826154 | NQO1 | 1653 | -0.0056 | 0.0288 | 0.8449 |
| 16 | 68299549 | rs12933210 | NQO1 | 1669 | -0.0314 | 0.0206 | 0.1279 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 16 | 68321913 | rs1469908 | NQO1 | 1695 | -0.0472 | 0.0206 | 0.0223 |
| 16 | 68329211 | hCV26055094 | NQO1 | 1677 | 0.0310 | 0.0200 | 0.1220 |
| 16 | 68333878 | rs1075935 | NQO1 | 1658 | -0.0377 | 0.0527 | 0.4754 |
| 16 | 80622339 | rs4291899 | HSD17B2 | 1707 | -0.0508 | 0.0349 | 0.1457 |
| 16 | 80632818 | rs11648233 | HSD17B2 | 1694 | 0.0043 | 0.0209 | 0.8361 |
| 16 | 80654301 | rs11642323 | HSD17B2 | 1700 | 0.0131 | 0.0211 | 0.5359 |
| 16 | 80670972 | rs2042429 | HSD17B2 | 1624 | 0.0083 | 0.0207 | 0.6889 |
| 16 | 80672242 | rs2966244 | HSD17B2 | 1710 | 0.0853 | 0.0711 | 0.2308 |
| 16 | 80683051 | rs1017243 | HSD17B2 | 1686 | -0.0063 | 0.0208 | 0.7635 |
| 16 | 80690493 | rs996752 | HSD17B2 | 1620 | -0.0153 | 0.0212 | 0.4700 |
| 16 | 80693012 | rs10514525 | HSD17B2 | 1680 | 0.0168 | 0.0204 | 0.4105 |
| 16 | 80693755 | rs1364284 | HSD17B2 | 1686 | 0.0020 | 0.0312 | 0.9497 |
| 16 | 80700383 | rs7200459 | HSD17B2 | 1699 | -0.0013 | 0.0373 | 0.9724 |
| 16 | 80703426 | rs12597465 | HSD17B2 | 1693 | 0.0152 | 0.0205 | 0.4605 |
| 17 | 37958089 | rs2830 | HSD17B1 | 1578 | -0.0031 | 0.0214 | 0.8841 |
| 17 | 37964418 | rs2854977 | HSD17B1 | 1688 | 0.0360 | 0.0434 | 0.4076 |
| 17 | 37974568 | rs650558 | HSD17B1 | 1700 | -0.0176 | 0.0231 | 0.4452 |
| 17 | 37975688 | rs1474040 | HSD17B1 | 1682 | 0.0039 | 0.0252 | 0.8763 |
| 17 | 37981755 | rs878291 | HSD17B1 | 1689 | -0.0073 | 0.0205 | 0.7230 |
| 17 | 37988603 | rs9903251 | HSD17B1 | 1691 | 0.0275 | 0.0218 | 0.2070 |
| 22 | 18291831 | rs12484658 | COMT | 1701 | 0.0322 | 0.0396 | 0.4167 |
| 22 | 18314051 | rs174675 | COMT | 1674 | -0.0065 | 0.0229 | 0.7766 |
| 22 | 18317638 | rs5993883 | COMT | 1689 | -0.0046 | 0.0202 | 0.8199 |
| 22 | 18329644 | rs3810595 | COMT | 1664 | 0.0115 | 0.0209 | 0.5829 |
| 22 | 18331897 | rs4646315 | COMT | 1701 | 0.0113 | 0.0239 | 0.6348 |
| 22 | 18332561 | rs165774 | COMT | 1694 | -0.0032 | 0.0220 | 0.8861 |
| 22 | 18333176 | rs174696 | COMT | 1695 | -0.0128 | 0.0241 | 0.5952 |
| 22 | 18335157 | rs9306235 | COMT | 1703 | -0.0084 | 0.0408 | 0.8364 |
| 22 | 18349075 | rs2073747 | COMT | 1675 | 0.0015 | 0.0245 | 0.9525 |
| 22 | 18350502 | rs1990277 | COMT | 1692 | 0.0017 | 0.0207 | 0.9327 |

Chr: chromosome; SNP: single nucleotide polymorphism rsid; N: number of subjects; SE: standard error; P: P-value for 1 d.f. trend test

**Table S4. Genes containing polymorphisms within the estrogen metabolic pathway evaluated in relation to mammographic density.**

| *Gene* | *Full name* | *Locus* | *No. of SNPs* | *Reference* |
|---|---|---|---|---|
| AKR1C4 | Aldo-keto reductase family 1, member C4 | 10p15-14 | 11 | [12] |
| COMT | catechol-o-methyltransferase | 22q11.2 | 10 | [5, 10, 11, 14, 48] |
| CYP11A1 | cytochrome p450scc / cholesterol side chain cleavage | 15q23-24 | 6 | |
| CYP11B1 | cytochrome p450 / steroid 11 beta 1 hydroxylase | 8q21 | 2 | |
| CYP11B2 | cytochrome p450 / steroid 11 beta 2 hydroxylase | 8q21 | 4 | |
| CYP17A1 | cytochrome p450 / 17 alfa steroid hydroxylase | 10q24.3 | 5 | [10, 11, 13, 14] |
| CYP19A1 | cytochrome p450 / aromatase | 15q21.1 | 14 | [10, 49] |
| CYP1A1-2 | cytochrome p450 | 15q22-qter | 5 | [10, 11, 48] |
| CYP1B1 | cytochrome p450 | 2p22-21 | 5 | [10-12] |
| CYP3A4_5 | cytochrome p450 | 7q22.1 | 6 | |
| GSTP1 | Glutathione S transferase, pi | 11q13 | 1 | |
| HSD11B1 | 11 beta hydroxy steroid dehydrogenase 1 | 1q32-q41 | 9 | |
| HSD11B2 | 11 beta hydroxy steroid dehydrogenase 2 | 16q22 | 6 | |
| HSD17B1 | 17 beta hydroxy steroid dehydrogenase 1 | 17q12-21 | 6 | [14] |
| HSD17B2 | 17 beta hydroxy steroid dehydrogenase 2 | 16q24.1-24.2 | 11 | |
| HSD17B3 | 17 beta hydroxy steroid dehydrogenase 3 | 9q22 | 8 | |
| HSD17B4 | 17 beta hydroxy steroid dehydrogenase 4 | 5q2 | 8 | |
| HSD17B7 | 17 beta hydroxy steroid dehydrogenase 7 | 10p11.2 | 5 | |
| HSD17B8 | 17 beta hydroxy steroid dehydrogenase 8 | 6p21.3 | 4 | |
| HSD3B1 | 3 beta hydroxy steroid dehydrogenase 1 | 1p13.1 | 7 | [5, 14] |
| NAT1 | N-acetyltransferase 1 | 8p23.1-21.3 | 7 | |
| NAT2 | N-acetyltransferase 2 | 8p23.1-21.3 | 7 | |
| NQO1 | NAD(P)H dehydrogenase, quinone 1 | 16q22.1 | 7 | |
| SOD2 | Superoxide dismutase 2, mitochondrial | 6q25.3 | 5 | |
| SRD5A1 | steroid 5 alpha reductase 1 | 5p15 | 5 | |
| SRD5A2 | steroid 5 alpha reductase 2 | 2p23 | 7 | |
| STE (SULT1E1) | estrogen/aryl sulfotransferase | 4q13.1 | 7 | |
| STS | arylsulfatase C/steroid sulfatase | Xp22.32 | 9 | |
| SULT1A1-2 | sulfotransferase family 1A, phenol-preferring member 1 | 16p12.1-11.2 | 6 | |
| SULT2A1 | dehydroepiandrosterone sulfotransferase | 19q13.3 | 8 | |
| SULT2B1 | sulfotransferase family 2B member 1 | 19q13.3 | 12 | |
| UGT1A1-9 | Uridine diphosphate glucuronosyltransferase 1 family, polypeptide A9 | 2q37 | 12 | [10] |
| UGT2B11 | Uridine diphosphate glucuronosyltransferase 2 family, member B 11 | 4q13.2 | 7 | |
| UGT2B4 | Uridine diphosphate glucuronosyltransferase 2 family, member B 4 | 4q13 | 7 | [12] |

**Table S5. Validation results of significantly associated SNPs in the Nurses' Health Study (NHS) and the Mayo Clinic Breast Cancer Study (MBCS)**

| SNP | Gene | Study | Minor allele | MAF | N | Genotyped* | R2 | Coefficient | SE | P |
|---|---|---|---|---|---|---|---|---|---|---|
| rs11638442 | CYP11A1 | NHS | C | 0.40 | 1590 | No | 1.00 | -0.030 | 0.020 | 0.12 |
| rs11638442 | CYP11A1 | MBCS | C | 0.40 | 783 | Yes | 1.00 | -0.005 | 0.034 | 0.88 |
| rs16968478 | CYP11A1 | NHS | G | 0.19 | 1590 | Yes | 1.00 | -0.036 | 0.026 | 0.16 |
| rs2279357 | CYP11A1 | NHS | T | 0.29 | 1590 | Yes | 1.00 | -0.0002 | 0.022 | 0.99 |
| rs2959003 | CYP11A1 | NHS | A | 0.33 | 1590 | No | 1.00 | -0.020 | 0.021 | 0.35 |
| rs2959008 | CYP11A1 | NHS | A | 0.33 | 1590 | No | 1.00 | -0.021 | 0.021 | 0.32 |
| rs2066485 | HSD17B3 | NHS | C | 0.15 | 1590 | No | 1.00 | -0.018 | 0.028 | 0.52 |
| rs7039978 | HSD17B3 | NHS | A | 0.48 | 1590 | No | 0.97 | 0.029 | 0.020 | 0.15 |
| rs1469908 | NQO1 | NHS | C | 0.43 | 1590 | No | 0.99 | -0.002 | 0.020 | 0.94 |
| rs17268974 | STS | NHS | A | 0.22 | 1590 | No | 0.88 | 0.0003 | 0.026 | 0.99 |
| rs2270112 | STS | NHS | C | 0.32 | 1590 | No | 1.00 | 0.020 | 0.021 | 0.35 |
| rs707762 | STS | NHS | A | 0.40 | 1590 | No | 1.00 | 0.021 | 0.020 | 0.31 |

* SNPs not genotyped were imputed using MACH based on HapMap Phase II (release 21a).

SNP: single nucleotide polymorphism rsid; NHS: Nurses' Healthy study; MBCS: The Mayo Clinic Breast Cancer Study; MAF: minor allele frequency; N: number of subjects; SE: standard error; P: P-value for 1 d.f. trend test

**Additional methods on SNP selection from:**

**Low YL, Yuqing L, Humphreys K, Thalamuthu A, Li Y, Darabi H, Wedrén S, Bonnard C, Czene K, Iles M *et al*:** Multi-variant Pathway Association Analysis Reveals the Importance of Genetic Determinants of Estrogen Metabolism in Breast and Endometrial Cancer Susceptibility**. *(submitted)* 2009.**

---

***DNA Isolation***

DNA was extracted from 4 ml of whole blood using the QIAamp DNA Blood Maxi Kit (Qiagen) according to manufacturer's instructions and non-malignant cells in paraffin-embedded tissue using a standard phenol/chloroform/isoamyl alcohol protocol [1].

***Gene and SNP Selection***

We selected 35 genes that code the enzymes involved in estradiol or estrone metabolism and are expressed in the breast. We selected 1007 single nucleotide polymorphisms (SNPs) in these genes and their 30kb flanking sequences from the dbSNP (build 124) and Celera databases, aiming for a marker density of at least one SNP per 5kb (Supplement Table 1). These SNPs were genotyped in 92 Swedish control samples to assess linkage disequilibrium pattern and coverage. Haplotypes were reconstructed using the PLEM algorithm [2] implemented in the *tagSNPs* program [3]. A subset of SNPs, tagSNPs, were selected based on the $R^2$ coefficient, which quantifies how well the tagSNP haplotypes predict the genotype or the number of copies of haplotypes an individual carries. We chose tagSNPs so that common SNP genotypes (minor allele frequency $\geq$0.03) and common haplotypes (frequency $\geq$0.03) were predicted with $R^2 \geq 0.8$ [4]. To evaluate our tagSNPs' performance in capturing unobserved SNPs within the genes and to assess whether we needed a denser set of markers, we performed a SNP-dropping analysis [5,6]. In brief, each of the genotyped SNPs was dropped in turn and tagSNPs were selected from the remaining SNPs so that their haplotypes predicted the remaining SNPs with an $R^2$ value of 0.85. We then estimated how well the tagSNP haplotypes of the remaining SNPs predicted the dropped SNP, an evaluation that can provide an unbiased and accurate estimate of tagSNP performance [5,6]. Overall, we selected and genotyped 302 tagSNPs from the 35 genes in all the Swedish cases and controls.

***Genotyping***

Genotyping was performed using the primer extension-based assay from Sequenom (San Diego, California) according to manufacturers' instructions. DNA samples were randomly assigned to the plates carrying positive and negative controls, and all genotyping results were generated and checked by laboratory staff unaware of case-control status. SNPs with a call rate < 85%, minor allele frequency < 1% or out of Hardy-Weinberg Equilibrium (p<0.05/252) were excluded from further analysis. Overall, 239 tagSNPs from the 34 genes were successfully genotyped and used in statistical analysis. The genotype concordance was >99%, suggesting high genotyping accuracy.

## References

1. Iles MM: **Obtaining unbiased estimates of tagging SNP performance**. *Ann Hum Genet* 2006, **70**(Pt 2):254-261.
2. Qin ZS, Niu T, Liu JS: **Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms**. *Am J Hum Genet* 2002, **71**(5):1242-1247.
3. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC: **Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study**. *Hum Hered* 2003, **55**(1):27-36.
4. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al*: **The structure of haplotype blocks in the human genome**. *Science* 2002, **296**(5576):2225-2229.
5. Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB: **Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping**. *Am J Hum Genet* 2003, **73**(3):551-565.
6. Iles MM: **Obtaining unbiased estimates of tagging SNP performance**. *Ann Hum Genet* 2006, **70**(2):254-261.

# IV

Li *et al. Breast Cancer Research* 2010, **12**:R23
http://breast-cancer-research.com/2010/12/2/R23

Breast Cancer RESEARCH

RESEARCH ARTICLE Open Access

# Effects of childhood body size on breast cancer tumour characteristics

Jingmei Li*[1,2], Keith Humphreys[1], Louise Eriksson[1], Kamila Czene[1], Jianjun Liu[2] and Per Hall[1]

## Abstract

**Introduction:** Although a role of childhood body size in postmenopausal breast cancer risk has been established, less is known about its influence on tumour characteristics.

**Methods:** We studied the relationships between childhood body size and tumour characteristics in a Swedish population-based case-control study consisting of 2,818 breast cancer cases and 3,111 controls. Our classification of childhood body size was derived from a nine-level somatotype. Relative risks were estimated by odds ratios with 95% confidence intervals, derived from fitting unconditional logistic regression models. Association between somatotype at age 7 and tumour characteristics were evaluated in a case-only analysis where *P* values for heterogeneity were obtained by performing one degree of freedom trend tests.

**Results:** A large somatotype at age 7 was found to be associated with decreased postmenopausal breast cancer risk. Although strongly associated with other risk factors such as age of menarche, adult body mass index and mammographic density, somatotype at age 7 remained a significant protective factor (odds ratio (OR) comparing large to lean somatotype at age 7 = 0.73, 95% confidence interval (CI) = 0.58-0.91, *P* trend = 0.004) after adjustment. The significant protective effect was observed within all subgroups defined by estrogen receptor (ER) and progesterone receptor (PR) status, with a stronger effect for ER-negative (0.40, 95% CI = 0.21-0.75, *P* trend = 0.002), than for ER-positive (0.80, 95% CI = 0.62-1.05, *P* trend = 0.062), tumours (*P* heterogeneity = 0.046). Somatotype at age 7 was not associated with tumour size, histology, grade or the presence or absence of metastatic nodes.

**Conclusions:** Greater body size at age 7 is associated with a decreased risk of postmenopausal breast cancer, and the associated protective effect is stronger for the ER-negative breast cancer subtype than for the ER-positive subtype.

## Introduction

There is considerable evidence that childhood anthropometric measurements are associated with postmenopausal breast cancer risk. It has been consistently shown that variables that approximate body shape and size early in life are inversely associated with breast cancer risk in adulthood. For example, a study conducted in 1998 on the same data set as used in the current study [1] reported that a larger somatotype at age seven years was associated with a lower postmenopausal breast cancer risk. Likewise, Hilakivi-Clarke and colleagues [2] found that a shorter height and higher body mass in girls from age 7 to 15 years were associated with a decreased incidence of breast cancer. Berkey and colleagues [3] also found extremely lean body mass at age 10 years to be associated with elevated breast cancer risk. In another study performed in 141,393 Danish girls, a high childhood body mass index (BMI) at age 14 years was shown to be protective against breast cancer later on in life [4]. In addition, a study performed on the large Nurses' Health Study dataset concluded that average body fatness between the ages of 5 and 10 years are inversely associated with mammographic density [5], which is generally considered to be an intermediate phenotype of breast cancer [6].

Although a role of childhood body size in adult breast cancer risk has been established, less is known about its influence on tumour characteristics. One study by Bardia and colleagues [7] looked into the risk of developing postmenopausal breast cancer stratified by estrogen receptor (ER) and progesterone receptor (PR) subtypes and reported that an increase in weight at age 12 years was

* Correspondence: jingmei.li@ki.se
[1] Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Box 281, 171 77 Stockholm, Sweden
Full list of author information is available at the end of the article

associated with a decrease in adult breast cancer risk, with the most pronounced effects exhibited by ER-positive/PR-negative tumours. No significant heterogeneity, however, was observed between the tumour subtypes studied. To our knowledge, no other study has been conducted to assess whether pre-/peri-pubertal measurements of body size can also influence tumour characteristics. We thus followed up on the work of Bardia and colleagues and in the present study examined the relations between childhood body size to address if the far-reaching effects of childhood body size have any influence on tumour characteristics in adult cancers.

## Materials and methods

### Subjects

The subjects included in the current study are drawn from a population-based case-control study of postmenopausal breast cancer in Swedish-born women aged 50 to 74 years at the time of enrolment, which was between 1 October, 1993 and 31 March, 1995. Controls were randomly selected from the Swedish registry and frequency matched to the expected age distribution of the cases. Details on data collection and subjects have been described previously [1]. The final study group included 2,818 cases and 3,111 controls. Approval of the study was given by the ethical review board at the Karolinska Institutet (Stockholm, Sweden) and six other ethical review boards in the respective regions from which the subjects were based.

### Data collection and classification

With the exception of clinical data on tumour characteristics and mammographic density, all other covariate data were derived from the parent case-control study. Anthropometric measurements at age seven years and one year prior to enrolment were collected by means of a nine-level somatotype (Figure 1) featured in the study questionnaire, and the validity of this measurement method has been previously described [1]. These pictograms have been validated against BMI within a cohort of 100 Caucasian women from middle-class communities with an average age of 73.1 years [8]. In a population-based validation study, 111 Swedish women aged 51 to 66 years were found to have a correlation coefficient between BMI from school records and adult report of somatotype at age seven years of 0.6 [1]. The somatotypes were subsequently grouped as lean (S1 to S2), medium (S3 to S4) and large (S5 to S9) prior to analysis. Other covariate data that was collected using the self-reported study questionnaire and examined in this study include age of menarche (continuous, in years), parity (continuous, number of live births), history of benign breast disease (binary, never/ever), BMI (continuous, in kg/m$^2$), history of hormone replacement therapy (HRT) (binary, never/ever), and family history of breast cancer (binary, no/yes). Age at menopause (continuous, in years) was also derived from information collected in the study questionnaire and the definition used in this study has been previously described [1]. It is defined as the age at the last menstrual period or the age at bilateral oophorectomy, if one year or more prior to data collection. Women who have had a hysterectomy, or who have not ceased menstruation due to HRT, or with missing information on age at menopause were considered to be postmenopausal if the age reported at time of questionnaire was equal to or above the 90th percentile of age at natural menopause of study subjects (current smokers: 54 years old; nonsmokers: 55 years old, independent of case/control status). Subjects classified as postmenopausal in this manner were assigned an age at menopause according to their current smoking status and the mean ages at natural menopause in our data. Otherwise, women were considered to be premenopausal and were excluded.

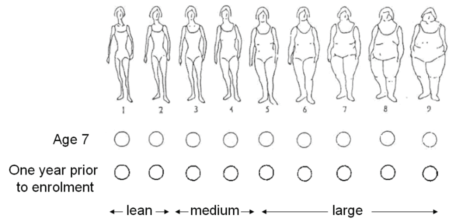


**Figure 1 Nine-level somatotype pictogram**.

Information regarding the retrieval of tumour characteristics from the medical records of all participants from surgical and oncological units throughout Sweden have been presented in detail elsewhere [9,10]. The tumour characteristics in the present study included tumour size (categorical, groups in cm), grade (categorical, classified according to the Nottingham histological grade or Bloom-Richardson scale), as well as ER and PR status (binary, absent/present).

The process of collecting mammographic density data in this study has been described previously [11]. Film mammograms of the medio-lateral oblique view were digitised using an Array 2905HD Laser Film Digitizer (Array Corporation, Tokyo, Japan), which covers a range of 0 to 4.7 optical density. For controls, breast side was randomized. For cases, the side contralateral to the tumour was used. The density resolution was set at 12-bit spatial resolution. The Cumulus software used for the computer-assisted thresholding was developed at the University of Toronto [12]. For each image, a trained observer (LE) set the appropriate gray-scale threshold levels defining the edge of the breast and distinguishing dense from non-dense tissue. The software calculated the total number of pixels within the entire region of interest and within the region identified as dense. These values were used to calculate the percentage of the breast area that is dense. A random 10% of the images were included as replicates to assess the intra-observer reliability, which was high with a Spearman rank correlation coefficient of 0.95. However, as not all women attended mammographic screenings, and some mammograms were missing, such information was available for only a subset of the subjects (n = 3232, 54.5%).

### Statistical analyses

The distribution of baseline characteristics of known breast cancer risk factors were summarised as means and standard deviations or proportions. Odds ratio (OR) estimates with corresponding 95% confidence intervals (CI) were computed by fitting unconditional logistic regression models with breast cancer risk status as the response variable, adjusting for age.

To identify potential confounders of the association between somatotype at age seven years and breast cancer risk, linear/logistic regression models were fitted for either continuous (age of menarche, age of menopause, parity, BMI, and mammographic density) or binary (benign breast disease and HRT) outcomes including only controls in the analysis. Somatotype at age seven years was treated as a categorical (three-level) independent variable. Proportional odds logistic regression was used in situations where the outcome variable was ordinal (somatotypes at age seven years and one year prior to enrolment) from which cumulative OR esimates with corresponding 95% CIs were computed. Covariates were considered potential confounders if there was *a priori* evidence in the published literature of the factor being associated with both childhood body size and breast cancer risk, or if the factor was significantly associated at the 5% level with both somatotype at age seven years and breast cancer risk. Those covariates that, when added to the model, changed the coefficient by more than 10%, were considered confounders and adjusted for in the multivariate analysis. The final variables in the multivariate logistic regression model examining breast cancer risk overall, and stratified by ER and PR tumour subtypes, included age, age at menarche, benign breast disease, and BMI one year prior to enrolment (recent BMI). Adjustment for other variables did not influence the somatotype risk estimates. Mammographic density was also identified as a confounder. However, as mammographic density data are only available for a subset of the subjects, this variable was accounted for together with the other risk factors in a separate model. Women with and without mammographic density information were not found to differ significantly at the 5% level for the covariates included in the analysis models (data not shown).

Associations between somatotype at age seven years and tumour characteristics were evaluated in a case-only analysis, by fitting ordinal regression models treating tumour characteristics as dependent variables, with somatotype at age seven years included as a covariate. *P* values for heterogeneity were obtained by performing one degree of freedom trend tests. As there exists prior evidence that certain tumour characteristics such as ER status are associated with age at diagnosis [13], and that somatotype at age seven years is significantly associated with age of diagnosis at the 5% level (regression coefficient for age in years of -0.91 with corresponding 95% CI of -1.32 to -0.50), every model fitted in the case-only analysis was also adjusted for age at diagnosis. All analyses were performed using the statistical software R for Windows version 2.8.0 (R Development Core Team, Vienna, Austria) [14]. The level of significance was set at 5%. All statistical tests were two-sided.

## Results

Table S1 in Additional file 1 describes the characteristics of study subjects with respect to several breast cancer risk factors. Age of menarche was weakly but positively associated with the disease (OR per year increase in age of menarche = 0.96, 95% CI = 0.93 to 1.00, *P* = 0.057), a result consistent with the literature [4]. Family history, age at menopause, parity, age of first birth, benign breast disease, mammographic density, recent BMI and use of HRT were strongly significant for breast cancer risk with effects in a direction consistent with those estimated in other epidemiological studies. The first association analyses we performed between somatotypes at different ages and breast cancer risk were adjusted for age at enrolment only. Among the different measurements of somatotypes, only the time point at age seven years was found to affect breast cancer risk (OR per increase in somatotype class = 0.87, 95% CI = 0.8 to 0.95, *P* = 0.001). A larger proportion of cases than controls had a leaner body shape at age seven years. Despite somatotype one year prior to enrolment having a high correlation to recent BMI (Spearman correlation coefficient: 0.760, data not shown), it was not found to be significantly associated with breast cancer (OR per increase in somatotype class = 1.04, 95% CI = 0.94 to 1.15, *P* = 0.160).

To identify potential confounders of the association between somatotype at age seven years and breast cancer risk, we assessed whether other established risk factors for breast cancer are associated with somatotype at age seven years. An increase in childhood body size was found to exhibit strong inverse associations with age of menarche (OR comparing large to lean somatotype at age seven years = 0.61, 95% CI = 0.50 to 0.76, *P* trend < 0.0001), benign breast disease (0.47, 95% CI = 0.25 to 0.89, *P* trend = 0.006), and mammographic density (0.61,

**Table 1: Associations of somatotype at age seven years with other breast cancer risk factors (controls only)**

| Risk factor (dependent variable) | Somatotype (independent variable) | n | OR | 95% CI | | *P* trend* |
|---|---|---|---|---|---|---|
| Age of menarche (years) | Lean | 1456 | 1.00 | reference | | <0.0001 |
| | Medium | 669 | 0.72 | 0.64 | 0.82 | |
| | Large | 187 | 0.61 | 0.50 | 0.76 | |
| Age of menopause (years) | Lean | 1572 | 1.00 | reference | | 0.697 |
| | Medium | 736 | 1.19 | 0.85 | 1.68 | |
| | Large | 204 | 0.93 | 0.53 | 1.65 | |
| Parity (Number of live births) | Lean | 1578 | 1.00 | reference | | 0.217 |
| | Medium | 745 | 0.93 | 0.83 | 1.05 | |
| | Large | 207 | 0.93 | 0.76 | 1.13 | |
| Benign breast disease | Lean | 1578 | 1.00 | reference | | 0.006 |
| | Medium | 745 | 0.76 | 0.56 | 1.03 | |
| | Large | 207 | 0.47 | 0.25 | 0.89 | |
| Somatotype one year prior to enrolment | Lean | 1571 | 1.00 | reference | | <0.0001 |
| | Medium | 739 | 1.72 | 1.44 | 2.05 | |
| | Large | 206 | 2.33 | 1.70 | 3.18 | |
| BMI (kg/m$^2$) | Lean | 1562 | 1.00 | reference | | <0.0001 |
| | Medium | 742 | 1.85 | 1.30 | 2.65 | |
| | Large | 205 | 2.66 | 1.47 | 4.83 | |
| Percent mammographic density (%)† | Lean | 862 | 1.00 | reference | | 0.001 |
| | Medium | 428 | 0.72 | 0.58 | 0.91 | |
| | Large | 108 | 0.61 | 0.41 | 0.90 | |
| HRT | Lean | 1569 | 1.00 | reference | | 0.868 |
| | Medium | 739 | 0.99 | 0.83 | 1.18 | |
| | Large | 206 | 0.98 | 0.73 | 1.32 | |
| | Other independent variables | | | | | |
| Birthweight (g) on somatotype at age 7 | ≤2500 | 49 | 1.00 | reference | | 0.014 |
| | 2500-3000 | 229 | 1.18 | 0.61 | 2.29 | |
| | 3000-3500 | 470 | 1.29 | 0.68 | 2.43 | |
| | 3500-4000 | 397 | 1.44 | 0.76 | 2.73 | |
| | >4000 | 135 | 1.89 | 0.95 | 3.76 | |
| Family history on somatotype at age 7 | No | 2258 | 1.00 | reference | | 0.485 |
| | Yes | 227 | 1.10 | 0.84 | 1.44 | |

* Based on Wald tests for regression coefficients in continuous, ordinal or logistic regression models (see statistical analyses section). All regression models were adjusted for age at enrolment. † Subset with phenotypic data. BMI, body mass index; CI, confidence interval; HRT, hormone replacement therapy; OR, odds ratio.

95% CI = 0.41 to 0.90, *P* trend = 0.001; Table 1). Associations in the opposite direction were found for proxy measures of physique at other time points, such as birth weight (OR comparing birthweight >4000 g to ≤2500 g = 1.89, 95% CI = 0.95 to 3.76, *P* trend = 0.014), somatotype one year prior to enrolment (OR comparing large to lean somatotype at age seven years = 2.33, 95% CI = 1.70 to 3.18, *P* trend < 0.0001) and recent BMI (2.66, 95% CI = 1.47 to 4.83, *P* trend < 0.0001). No evidence of association was found between age of menopause and somatotype at age seven years or between family history and somatotype at age seven years. Parity and HRT were found to be independent of somatotype at age seven years (0.93, 95% CI = 0.76 to 1.13, *P* trend = 0.217 and 0.98, 95% CI = 0.73 to 1.32, *P* trend = 0.868, respectively).

After adjustment of known breast cancer predictors and other associated risk factors, the inverse association of somatotype at age seven years with breast cancer remained highly significant (Table 2; OR comparing large to lean somatotype at age seven years = 0.73, 95% CI = 0.58 to 0.91, *P* trend = 0.004). The protective effect of a larger somatotype was found to be significant (*P* trend < 0.05) for ER-negative, PR-positive and PR-negative subtypes and marginally significant (*P* trend = 0.062) for the ER-positive subtype. Within the group consisting of large somatotypes, the most prominent effects were shown in ER-negative (OR comparing large to lean somatotype at age seven years = 0.40, 95% CI = 0.21 to 0.75, *P* trend = 0.002) and PR-negative (0.63, 95% CI = 0.40 to 0.99, *P* trend = 0.028) tumours. The point estimates changed very little before and after additional adjustment for mammographic density as a continuous variable [see Table S2 in Additional file 2], using a subset of the data with this information available (n = 3232).

We next assessed the effects of childhood body size on tumour characteristics (ER status, PR status, tumour size, grade, histology, and absence/presence of metastatic nodes) by fitting binary/ordinal logistic regression models, adjusting for age at diagnosis in years as a confounder. We established that the protective effect of somatotype at age seven years was significantly stronger for ER-negative disease than for ER-positive disease (P heterogeneity = 0.046; Table 3). When comparing between two extreme groups, women with a larger body size at age seven years were 1.71 times (95% CI = 0.96 to 3.06) more likely to get ER-positive than ER-negative disease after menopause. Although the estimated trend suggests that women with the same physique are more likely to get the PR-positive disease in adulthood, the difference between the two tumour subtypes was not significant (P heterogeneity = 0.283). The point estimates for tumour size, histology, grade, or the presence or absence of metastatic nodes did not vary much before and after adjustment for age of diagnosis as a continuous variable.

## Discussion

Our first main finding was that a large somatotype at age seven years was associated with a decreased risk of postmenopausal breast cancer. Although strongly associated with other risk factors such as age of menarche, adult BMI and mammographic density, somatotype at age seven years remained a significant protective factor (OR comparing large to lean somatotype at age seven years = 0.73, 95% CI = 0.58 to 0.91, *P* trend = 0.004) after adjustment for these other risk factors. Our second and most novel finding was of a significant protective effect of somatotype at age seven years regardless of receptor status, but with a stronger effect for ER-negative (0.40, 95% CI = 0.21 to 0.75, *P* trend = 0.002), than for ER-positive (0.80, 95% CI = 0.62 to 1.05, *P* trend = 0.062), tumours (*P* heterogeneity = 0.046).

Our findings regarding the protective effects of childhood body size for adult breast cancer are consistent with previous studies [3-5]. Associations with other breast cancer risk factors were also in the same direction as found in other epidemiological studies. Several studies have found birth weight and gain in BMI in early childhood to predict adult lean mass, while adult adiposity has been attributed to weight gain in late childhood and adolescence [15-19]. Similarly, anthropometric measurements at other time points (birth weight, and somatotype one year prior to enrolment) in our data were found to be positively associated with somatotype at age seven years. The adverse effects of birth weight and adult body mass on postmenopausal breast cancer risk may be explained by a surplus of estrogen exposure from either the uterine environment or excess adipose tissue [4,20]. However, studies performed on children have not consistently found an association between obesity and circulating estradiol levels [21,22], thus it is unclear what mechanisms drive the associated decrease in risk during the pre-/peri-puberty window.

Strong inverse relationships found between childhood body size, age of menarche, benign breast disease, and mammographic density were in line with other reports in the literature. Baer and colleagues [23] found a large childhood body size to be associated with a decrease in risk of benign breast disease. Age of menarche is often considered along with age of menopause and other hormonal risk factors for a woman's cumulative exposure to estrogen [24,25]. An earlier age of menarche is associated with an increased risk of breast cancer. On the other hand, a larger childhood somatotype, which is associated with decreased breast cancer risk, is also associated with an earlier age of menarche. As age of menarche is an established but weak predictor of breast cancer risk, its pronounced inverse relationship with childhood body size when seen in the context of breast cancer risk seems to be counterintuitive [26,27].

Mammographic density has also been found by others to be associated with childhood body mass [5]. Estrogen is produced by adipose tissue in the body. A higher BMI is thus correlated with higher endogenous estrogen levels. In a murine study, exposure to estrogen prior to puberty led to a decrease in radiologically dense tissue and an increase in the number of radiolucent structures [28], which may be analogous to a lower mammographic density in humans. In agreement, McCormack and col-

**Table 2: Multivariate-adjusted OR estimates and corresponding 95% CIs of postmenopausal breast cancer for somatotype at age seven years, overall and stratified by breast cancer tumour subtype based on ER and PR status**

| Type of breast cancer | Somatotype | All subjects | | | | |
|---|---|---|---|---|---|---|
| | | Cases | OR | 95% CI | | *P* trend* |
| All data | Lean | 1784 | 1.00 | reference | | 0.004 |
| | Medium | 757 | 0.90 | 0.79 | 1.02 | |
| | Large | 173 | 0.73 | 0.58 | 0.91 | |
| ER positive | Lean | 963 | 1.00 | reference | | 0.062 |
| | Medium | 408 | 0.91 | 0.78 | 1.06 | |
| | Large | 98 | 0.80 | 0.62 | 1.05 | |
| ER negative | Lean | 219 | 1.00 | reference | | 0.002 |
| | Medium | 81 | 0.77 | 0.58 | 1.03 | |
| | Large | 14 | 0.40 | 0.21 | 0.75 | |
| PR positive | Lean | 841 | 1.00 | reference | | 0.027 |
| | Medium | 354 | 0.89 | 0.75 | 1.04 | |
| | Large | 83 | 0.76 | 0.57 | 1.00 | |
| PR negative | Lean | 320 | 1.00 | reference | | 0.028 |
| | Medium | 126 | 0.86 | 0.68 | 1.08 | |
| | Large | 25 | 0.63 | 0.40 | 0.99 | |

* Logistic regression models were used, accounting for age, age at menarche, benign breast disease and recent body mass index. CI, confidence interval; ER, estrogen receptor; OR, odds ratio; PR, progesterone receptor.

leagues [29] showed that high childhood BMI was associated with a lower Wolfe grade, and Samimi and colleagues [5] found that a rounder pre-pubertal body shape was predictive of lower mammographic density later in life.

The age-adjusted case-only comparison of our data reflected a significant difference in the effects of childhood body size on the two ER subtypes (*P* trend = 0.046), but not the PR subtypes. However, in lieu of the fact that PR is an estrogen-induced target gene, and that its presence could serve to indicate ER functional capacity and tumour differentiation state [30], we also conducted stratified analyses on PR subtypes. We found that the protective trend conferred by a larger childhood somatotype on postmenopausal breast cancer applies to all ER and PR tumour subtypes. Overall our results were consistent with Bardia and colleagues [7], although in that study the effects were only significant for ER-positive (0.80, 95% CI = 0.67 to 0.96) and PR-negative (0.62, 95% CI = 0.43 to 0.89) tumours (comparing women with above average weight at age 12 years to women with average weight at age 12 years). Although Bardia and colleagues observed a stronger protective effect in ER-negative tumours than in their ER-positive counterparts (in agreement with our finding) when comparing women with above average weight at age 12 years to women with average weight at age 12 years, the association they observed in this subgroup was not statistically significant (0.77, 95% CI = 0.5 to 1.19).

Hormonal exposure and mammographic density are established risk factors of breast cancer that have been suggested to be independent, operating through different pathways [31]. Adjustment for these factors and other traditional risk factors did not attenuate the negative association of childhood body size on breast cancer risk (OR comparing large to lean somatotype at age seven years = 0.73, 95% CI = 0.58 to 0.91, *P* trend = 0.004, for association, after adjustment), thus suggesting an independent underlying mechanism. We speculate that a possible mechanism driving the negative association with breast cancer risk could be epigenetic changes that occur during mammary development. Hilakivi-Clarke [32] summarised in a review several perspectives on special windows of mammary development. Mammary tissue is postulated to undergo epigenetic extensive modelling or re-modelling during different stages in life such as fetal development, puberty or pregnancy. Such epigenetic modification can persist into adulthood if taken place in mammary stem cells, uncommitted mammary myoepithelial or luminal progenitor cells and inherited by subse-

**Table 3: Relation of somatotype at age seven years to tumour-defined characteristics of breast cancer**

| Tumour characteristics | Categories | Somatotype at age seven years | | | *P* heterogeneity§ |
|---|---|---|---|---|---|
| | | S1-S2 | S3-S4 | S5-S9 | |
| Tumour size (cm)* | <1 | 300 | 138 | 39 | |
| | 1-2 | 752 | 299 | 70 | |
| | 2-3 | 366 | 152 | 31 | |
| | 3-4 | 116 | 66 | 14 | |
| | 4-5 | 52 | 16 | 3 | |
| | >=5 | 65 | 26 | 4 | |
| Cumulative OR (95% CI) | | 1.00 (ref.) | 1.00 (0.85-1.17) | 0.78 (0.58-1.05) | 0.255 |
| Cumulative OR (95% CI) § | | 1.00 (ref.) | 1.00 (0.85-1.18) | 0.78 (0.58-1.06) | 0.266 |
| Grade* | Low | 159 | 69 | 20 | |
| | Medium | 479 | 186 | 46 | |
| | High | 463 | 222 | 51 | |
| Cumulative OR (95% CI) | | 1.00 (ref.) | 1.15 (0.94-1.41) | 0.99 (0.69-1.43) | 0.443 |
| Cumulative OR (95% CI) § | | 1.00 (ref.) | 1.15 (0.93-1.41) | 0.99 (0.69-1.42) | 0.463 |
| Histology* | Ductal | 1350 | 570 | 137 | |
| | Lobular | 206 | 77 | 16 | |
| | All other | 92 | 37 | 7 | |
| Cumulative OR (95% CI) | | 1.00 (ref.) | 0.91 (0.72-1.15) | 0.76 (0.48-1.20) | 0.192 |
| Cumulative OR (95% CI) § | | 1.00 (ref.) | 0.92 (0.73-1.17) | 0.79 (0.50-1.25) | 0.265 |
| Metastatic nodes† | Absent | 1159 | 473 | 116 | |
| | Present | 513 | 227 | 46 | |
| OR (95% CI) | | 1.00 (ref.) | 1.08 (0.90-1.31) | 0.90 (0.63-1.28) | 0.923 |
| OR (95% CI) § | | 1.00 (ref.) | 1.07 (0.88-1.29) | 0.86 (0.60-1.23) | 0.878 |
| ER status† | Negative | 219 | 81 | 14 | |
| | Positive | 963 | 408 | 98 | |
| OR (95% CI) | | 1.00 (ref.) | 1.15 (0.87-1.52) | 1.59 (0.89-2.84) | 0.089 |
| OR (95% CI) § | | 1.00 (ref.) | 1.18 (0.89-1.56) | 1.71 (0.96-3.06) | **0.046** |
| PR status† | Negative | 320 | 126 | 25 | |
| | Positive | 841 | 354 | 83 | |
| OR (95% CI) | | 1.00 (ref.) | 1.07 (0.84-1.36) | 1.26 (0.79-2.01) | 0.307 |
| OR (95% CI) § | | 1.00 (ref.) | 1.07 (0.84-1.37) | 1.28 (0.80-2.03) | 0.283 |

*Proportional odds logistic regression models were used. † Logistic regression models were used. ‡ Derived from one degree of freedom trend tests. § Adjusted for age at diagnosis. CI, confidence interval; ER, estrogen receptor; OR, odds ratio; PR, progesterone receptor.

quent daughter cells [33]. Prepubertal exposure to estrogen has been shown to upregulate the expression of BRCA1, a well-known DNA repair gene [28]. Liu and colleagues [34] also demonstrated that BRCA1 is responsible for differentiating ER-negative stem/progenitor cells into ER-positive luminal cells. They also proposed that loss of expression of the DNA repair gene (BRCA1) may result in an accumulation of ER-negative stem cells with multiple genetic defects. Incidentally, loss of BRCA1 is frequently associated with ER-negative breast cancers [35]. The evidence for altered gene expression possibly caused by childhood body size helps to explain the general reduction in breast cancer risk overall. The apparent differential protection conferred to the ER-negative subtype could possibly be driven by the same underlying mechanism that operates through epigenetic modifications.

The strengths of our study include it being a population-based study, its large sample size and detailed information on many variables: anthropometric measures at different time points throughout life, mammographic density, reproductive and hormonal risk factors, and tumour characteristics. To our knowledge, this is the first study to consider the effects of somatotype at age seven years on adult breast cancer with the consideration of mammographic density, and also the first to examine its effects on tumour characteristics other than ER status.

A limitation of our study is that risk factor data were self-reported, and could thus be measured with error. Although two studies have demonstrated the validity of using the nine-level somatotype diagram for the long-term recall of childhood body size via high correlations with BMI at the same ages [8,36], it is noteworthy that in those studies no woman recalled their figure as larger than level seven in these studies, and that women with large body size were more likely to misreport their childhood somatotypes than women who were lean. However, any such measurement error is most likely to attenuate any association between childhood body size and breast cancer risk [37]. In addition, as the questionnaire study was conducted post-diagnosis of breast cancer, recall bias could have been introduced. Although the nine-level somatotype measure has not been validated specifically in a group of breast cancer cases, it is unlikely that childhood body size was differentially recalled by breast cancer cases and by controls.

## Conclusions

Our findings may have important implications. The effects of childhood body size on the different breast cancer subtypes are independent of other breast cancer risk factors, such as mammographic density and estrogen exposure. Given the strength of the associations, and the ease of retrieval of information on childhood somatotypes retrospectively from pictures early in life, childhood body size is potentially useful for building breast cancer risk or prognosis prediction models. It appears counter-intuitive that a large body size during childhood can reduce breast cancer risk or alter one's prognosis, because a large birth weight and a high adult BMI have been shown to otherwise elevate breast cancer risk. There remain unanswered questions on mechanisms driving this protective effect. Because body size and related hormonal exposures are modifiable risk factors, women might substantially decrease their risk of breast cancer, in particular the more aggressive ER-negative disease, by monitoring their nutrition and exogenous hormone intake at different points in life.

## Additional material

**Additional file 1: Table S1**. Descriptive characteristics of post-menopausal women.

**Additional file 2: Table S2**. Multivariate-adjusted odds ratio (OR) estimates and corresponding 95% confidence intervals (CIs) of postmenopausal breast cancer for somatotype at age seven years on a subset of women with mammographic density data; overall and stratified by breast cancer tumour subtype based on estrogen receptor (ER) and progesterone receptor (PR) status.

### Abbreviations

BMI: body mass index; CI: confidence interval; ER: estrogen receptor; HRT: hormone replacement therapy; OR: odds ratio; PR: progesterone receptor.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

JLi participated in the study design, carried out the analyses and drafted the manuscript. LE digitised and obtained readings for the mammograms. KH, KC, JLiu and PH participated in study design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

This work was supported by National Institutes of Health (RO1 CA58427); the Märit and Hans Rausing's Initiative against Breast Cancer; and the W81XWH-05-1-0314 Innovator Award, US Department of Defense Breast Cancer Research Program, Office of the Congressionally Directed Medical Research Programs. J Li is a recipient of the A*STAR Graduate Scholarship. KH was supported by the Swedish Research Council (523-2006-972). KC was financed by the Swedish Cancer Society (5128-B07-01PAF). The sponsors took no role in the study design, the collection or analysis of the data, the interpretation of the results, the preparation of the manuscript, or the decision to submit the manuscript for publication.

### Author Details

[1]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Box 281, 171 77 Stockholm, Sweden and [2]Human Genetics, Genome Institute of Singapore, 60 Biopolis Street, Singapore, 138672, Singapore

Received: 29 December 2009 Revised: 22 February 2010
Accepted: 15 April 2010 Published: 15 April 2010

### References

1. Magnusson C, Baron J, Persson I, Wolk A, Bergstrom R, Trichopoulos D, Adami HO: **Body size in different periods of life and breast cancer risk in post-menopausal women.** *Int J Cancer* 1998, **76**:29-34.
2. Hilakivi-Clarke L, Forsen T, Eriksson JG, Luoto R, Tuomilehto J, Osmond C, Barker DJ: **Tallness and overweight during childhood have opposing effects on breast cancer risk.** *Br J Cancer* 2001, **85**:1680-1684.

3. Berkey CS, Frazier AL, Gardner JD, Colditz GA: **Adolescence and breast carcinoma risk.** *Cancer* 1999, **85**:2400-2409.
4. Ahlgren M, Melbye M, Wohlfahrt J, Sorensen TI: **Growth patterns and the risk of breast cancer in women.** *N Engl J Med* 2004, **351**:1619-1626.
5. Samimi G, Colditz GA, Baer HJ, Tamimi RM: **Measures of energy balance and mammographic density in the Nurses' Health Study.** *Breast Cancer Res Treat* 2008, **109**:113-122.
6. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, Paterson AD: **Mammographic breast density as an intermediate phenotype for breast cancer.** *Lancet Oncol* 2005, **6**:798-808.
7. Bardia A, Vachon CM, Olson JE, Vierkant RA, Wang AH, Hartmann LC, Sellers TA, Cerhan JR: **Relative weight at age 12 and risk of postmenopausal breast cancer.** *Cancer Epidemiol Biomarkers Prev* 2008, **17**:374-378.
8. Must A, Willett WC, Dietz WH: **Remote recall of childhood height, weight, and body build by elderly subjects.** *Am J Epidemiol* 1993, **138**:56-64.
9. Rosenberg LU, Einarsdottir K, Friman EI, Wedren S, Dickman PW, Hall P, Magnusson C: **Risk factors for hormone receptor-defined breast cancer in postmenopausal women.** *Cancer Epidemiol Biomarkers Prev* 2006, **15**:2482-2488.
10. Orgeas CC, Hall P, Rosenberg LU, Czene K: **The influence of menstrual risk factors on tumor characteristics and survival in postmenopausal breast cancer.** *Breast Cancer Res* 2008, **10**:R107.
11. Tamimi RM, Eriksson L, Lagiou P, Czene K, Ekbom A, Hsieh CC, Adami HO, Trichopoulos D, Hall P: **Birth weight and mammographic density among postmenopausal women in Sweden.** *Int J Cancer* 2010, **126**:985-991.
12. Boyd NF, Stone J, Martin LJ, Jong R, Fishell E, Yaffe M, Hammond G, Minkin S: **The association of breast mitogens with mammographic densities.** *Br J Cancer* 2002, **87**:876-882.
13. Bentzon N, During M, Rasmussen BB, Mouridsen H, Kroman N: **Prognostic effect of estrogen receptor status across age in primary breast cancer.** *Int J Cancer* 2008, **122**:1089-1094.
14. R Development Core Team: **R. A language and environment for statistical computing.** Vienna, Austria: R Foundation for Statistical Computing; 2005.
15. Rogers I: **The influence of birthweight and intrauterine environment on adiposity and fat distribution in later life.** *Int J Obes Relat Metab Disord* 2003, **27**:755-777.
16. Sachdev HS, Fall CH, Osmond C, Lakshmy R, Dey Biswas SK, Leary SD, Reddy KS, Barker DJ, Bhargava SK: **Anthropometric indicators of body composition in young adults: relation to size at birth and serial measurements of body mass index in childhood in the New Delhi birth cohort.** *Am J Clin Nutr* 2005, **82**:456-466.
17. Sellers TA, Davis J, Cerhan JR, Vierkant RA, Olson JE, Pankratz VS, Potter JD, Folsom AR: **Interaction of waist/hip ratio and family history on the risk of hormone receptor-defined breast cancer in a prospective study of postmenopausal women.** *Am J Epidemiol* 2002, **155**:225-233.
18. Yliharsila H, Kajantie E, Osmond C, Forsen T, Barker DJ, Eriksson JG: **Birth size, adult body composition and muscle strength in later life.** *Int J Obes (Lond)* 2007, **31**:1392-1399.
19. Yliharsila H, Kajantie E, Osmond C, Forsen T, Barker DJ, Eriksson JG: **Body mass index during childhood and adult body composition in men and women aged 56-70 y.** *Am J Clin Nutr* 2008, **87**:1769-1775.
20. Friedenreich CM: **Review of anthropometric factors and breast cancer risk.** *Eur J Cancer Prev* 2001, **10**:15-32.
21. Klein KO, Larmore KA, de Lancey E, Brown JM, Considine RV, Hassink SG: **Effect of obesity on estradiol level, and its relationship to leptin, bone maturation, and bone mineral density in children.** *J Clin Endocrinol Metab* 1998, **83**:3469-3475.
22. Garnett SP, Hogler W, Blades B, Baur LA, Peat J, Lee J, Cowell CT: **Relation between hormones and body composition, including bone, in prepubertal children.** *Am J Clin Nutr* 2004, **80**:966-972.
23. Baer HJ, Schnitt SJ, Connolly JL, Byrne C, Cho E, Willett WC, Colditz GA: **Adolescent diet and incidence of proliferative benign breast disease.** *Cancer Epidemiol Biomarkers Prev* 2003, **12**:1159-1167.
24. Emaus A, Espetvedt S, Veierod MB, Ballard-Barbash R, Furberg AS, Ellison PT, Jasienska G, Hjartaker A, Thune I: **17-beta-estradiol in relation to age at menarche and adult obesity in premenopausal women.** *Hum Reprod* 2008, **23**:919-927.
25. Jansen SC, Temme EH, Schouten EG: **Lifetime estrogen exposure versus age at menopause as mortality predictor.** *Maturitas* 2002, **43**:105-112.
26. Cooper C, Kuh D, Egger P, Wadsworth M, Barker D: **Childhood growth and age at menarche.** *Br J Obstet Gynaecol* 1996, **103**:814-817.
27. Terry MB, Ferris JS, Tehranifar P, Wei Y, Flom JD: **Birth weight, postnatal growth, and age at menarche.** *Am J Epidemiol* 2009, **170**:72-79.
28. Cabanes A, Wang M, Olivo S, DeAssis S, Gustafsson JA, Khan G, Hilakivi-Clarke L: **Prepubertal estradiol and genistein exposures up-regulate BRCA1 mRNA and reduce mammary tumorigenesis.** *Carcinogenesis* 2004, **25**:741-748.
29. McCormack VA, dos Santos Silva I, De Stavola BL, Perry N, Vinnicombe S, Swerdlow AJ, Hardy R, Kuh D: **Life-course body size and perimenopausal mammographic parenchymal patterns in the MRC 1946 British birth cohort.** *Br J Cancer* 2003, **89**:852-859.
30. Hardy DB, Janowski BA, Chen CC, Mendelson CR: **Progesterone receptor inhibits aromatase and inflammatory response pathways in breast cancer cells via ligand-dependent and ligand-independent mechanisms.** *Mol Endocrinol* 2008, **22**:1812-1824.
31. Boyd NF, Martin LJ, Sun L, Guo H, Chiarelli A, Hislop G, Yaffe M, Minkin S: **Body size, mammographic density, and breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 2006, **15**:2086-2092.
32. Hilakivi-Clarke L: **Nutritional modulation of terminal end buds: its relevance to breast cancer prevention.** *Curr Cancer Drug Targets* 2007, **7**:465-474.
33. De Assis S, Hilakivi-Clarke L: **Timing of dietary estrogenic exposures and breast cancer risk.** *Ann N Y Acad Sci* 2006, **1089**:14-35.
34. Liu S, Ginestier C, Charafe-Jauffret E, Foco H, Kleer CG, Merajver SD, Dontu G, Wicha MS: **BRCA1 regulates human mammary stem/progenitor cell fate.** *Proc Natl Acad Sci USA* 2008, **105**:1680-1685.
35. Karp SE, Tonin PN, Begin LR, Martinez JJ, Zhang JC, Pollak MN, Foulkes WD: **Influence of BRCA1 mutations on nuclear grade and estrogen receptor status of breast carcinoma in Ashkenazi Jewish women.** *Cancer* 1997, **80**:435-441.
36. Must A, Phillips SM, Naumova EN, Blum M, Harris S, Dawson-Hughes B, Rand WM: **Recall of early menstrual history and menarcheal body size: after 30 years, how well do women remember?** *Am J Epidemiol* 2002, **155**:672-679.
37. Gunnell D, Berney L, Holland P, Maynard M, Blane D, Frankel S, Smith GD: **How accurately are height, weight and leg length reported by the elderly, and how closely are they related to measurements recorded in childhood?** *Int J Epidemiol* 2000, **29**:456-464.

doi: 10.1186/bcr2564
**Cite this article as:** Li *et al.*, Effects of childhood body size on breast cancer tumour characteristics *Breast Cancer Research* 2010, **12**:R23

**Institutionen för medicinsk epidemiologi och biostatistik**

# Genetic determinants of breast cancer risk

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Petrén, Nobels väg 12B

**Fredagen den 18 februari, 2011, kl 09.00**

av

**Jingmei Li**

*Huvudhandledare:*
Professor Per Hall
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

*Bihandledare:*
Professor Kamila Czene
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Docent Keith Humphreys
Karolinska Institutet
Institutionen för medicinsk epidemiologi och

Docent Jianjun Liu
Genome Institute of Singapore
Human Genetics

*Fakultetsopponent:*
Professor Kirsten B. Moysich
Roswell Park Cancer Institute
Department of Cancer Prevention and Control

*Betygsnämnd:*
Docent Thomas Hatschek
Karolinska Universitetssjukhuset
Onkologiska kliniken,
Radiumhemmet

Docent Patrik Magnusson
Karolinska Institutet
Institutionen för medicinsk epidemiologi och biostatistik

Professor Åke Borg
Lund University
Department of Oncology,
Clinical Sciences

**Stockholm 2011**

# ABSTRACT

The main purpose of this thesis was to identify genetic risk factors using both hypothesis-based and hypothesis-free approaches.

In an attempt to identify common disease susceptibility alleles for breast cancer, we started off with a hypothesis-free approach, and performed a combined analysis of three genome-wide association studies (GWAS), involving 2,702 women of European ancestry with invasive breast cancer and 5,726 controls.

As GWAS has been said to underperform for studying complex diseases such as breast cancer, we investigated to see if the variance explained by common variants could be increased by studying specific disease subtypes. Breast cancer may be characterized on the basis of whether estrogen receptors (ER) are expressed in the tumour cells. The two breast cancer tumour subtypes (ER-positive and ER-negative) are generally considered as biologically distinct diseases and have been associated with remarkably different gene expression profiles. ER status is important clinically, and is used both as a prognosticator and treatment predictor since it determines if a patient may benefit from anti-estrogen therapy. We thus performed an independent GWAS using a subset of ER-negative breast cancer cases and all of the controls from the initial genome-wide study, and, in addition, also evaluated whether the two cancer subtypes are fundamentally different on a germline level.

Besides hypothesis-free GWAS, we also conducted hypothesis-based analyses based on candidate pathways to identify common variants associated with breast cancer. Several studies have examined the effect of genetic variants in genes involved in the estrogen metabolic pathway on mammographic density, but the number of loci studied and the sample sizes evaluated have been small and pathways have not been evaluated comprehensively. We evaluated a total of 239 SNPs in 34 genes in the estrogen metabolic pathway in 1,731 Swedish women who participated in a breast cancer case-control study.

Slightly venturing outside the genetic scope of this thesis, we looked at a breast cancer risk factor - body size - that is associated with very different postmenopausal breast cancer risks at different time points in a woman's lifetime, namely, birth, childhood, and postmenopausal adult.

The significance of these studies will be apparent when, using the new genetic and epidemiological knowledge found, we are able to classify women according to high or low risk of breast cancer on the basis of genetic disposition or other breast cancer risk factors, so that appropriate interventions and disease management decisions may be made, to ultimately reduce incidence and mortality of breast cancer.

Keywords: Breast Neoplasms, Genetic Epidemiology, Genetic Susceptibility, Genetic Predisposition to Disease/genetics*, Case-Control Studies, Genetic Association Studies, Candidate Gene Analysis, Gene Discovery, Single Nucleotide Polymorphism, Risk Factors,, Estrogen Receptors, Mammography, Body Size

ISBN: 978-91-7457-179-0