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

Genetic Determinants of Postmenopausal Breast and Endometrial Cancer

Kristjana Einarsdóttir

Stockholm 2007

Cover figure: LD plot of 155 SNPs in the *ESR1* gene that were genotyped in 92 Swedish controls.

Published and printed by Universitetsservice US-AB
Nanna Svartz väg 4, 171 77 Solna, Sweden

ISBN 91-7357-037-0

I dedicate this work to my parents,
Einar Njálsson and Sigurbjörg Bjarnadóttir

And to my husband, Anthony S. Gunnell

ABSTRACT

Breast cancer is overall the most common cancer in women worldwide and endometrial cancer is the most common gynaecological cancer in the industrialized world. History of a first-degree relative with breast or endometrial cancer has been related to a twofold increase in risk of the respective diseases. Whilst genetic risk factors for endometrial cancer in general or for breast cancer in women not carrying any high-penetrance mutations are largely unknown, a polygenic model has been suggested to account for residual familial risk. This model anticipates small effects of common, low-penetrance genetic risk variants in combination with environmental influence. We thus set out to study common variation in key breast and endometrial cancer genes in relation to a) breast or endometrial cancer risk overall or in subgroups of environmental risk factors, b) the risk of tumour characteristics-defined breast cancer, or c) breast cancer death. In this population-based case-control study, we included 1579 breast cancer cases, 705 endometrial cancer cases and 1565 shared controls. All participants donated tissue or whole blood and provided detailed information about various lifestyle factors through questionnaires.

The *CYP17*, *ATM*, *CHEK2* and *ERBB2* genes have all been suggested to play a key role in cancer aetiology and progression. They are important candidate genes in breast and endometrial cancer aetiology specifically through their involvement in the estrogen metabolism pathway, DNA-damage response or cell proliferation. We genotyped common single nucleotide polymorphisms (SNPs) and rare variants in these genes in all cases and controls. Using regression models, we then assessed the effect of the variants and their haplotypes on cancer risk and survival.

We found that the rare *1100delC* deletion in *CHEK2* was more common in breast cancer cases than controls and increased breast cancer risk with an odds ratio of 2.26 (95% CI 0.99–5.15) for carriers versus non-carriers. Our results also indicated an increased risk of developing endometroid endometrial cancer for homozygous carriers of the rare allele (AA) of a tagSNP (rs4987886) in *CHEK2* (*P* = 0.005), when contrasted with GG carriers. In addition, we found a decreased endometrial cancer risk among non-smoking carriers of a haplotype in *ATM* (*P* = 0.0007) and among carriers of a haplotype in *CHEK2* who had experienced menopause below 49 years of age (*P* = 0.0009), compared to non-carriers of these haplotypes. We found no effect of genetic variation in *CYP17* on breast cancer risk regardless of histopathology or menopausal hormone use. The *ATM*, *CHEK2* or *ERBB2* genes did not appear to affect the risk of tumour characteristics-defined breast cancer or breast cancer death. We did not find any evidence supporting a role for the *ATM* and *ERBB2* genes in breast cancer aetiology, and the *ERBB2* gene also did not seem to have an effect on endometrial cancer risk.

Our estimate of the breast cancer risk related to the *CHEK2*1100delC* is in line with previous studies published in Northern European populations. Further studies regarding *CHEK2* or *ATM* in relation to endometrial cancer risk are however required for corroboration since our results became statistically non-significant after multiple testing adjustment.

LIST OF PUBLICATIONS

This thesis is based on the following papers:

I. Einarsdóttir K, Rylander-Rudqvist T, Humphreys K, Ahlberg S, Jonasdottir G, Weiderpass E, Chia KS, Ingelman-Sundberg M, Persson I, Liu J, Hall P, Wedrén S.
CYP17 gene polymorphism in relation to breast cancer risk: a case-control study.
Breast Cancer Research 2005; 7(6):R890-R896.

II. Einarsdóttir K, Humphreys K, Bonnard C, Palmgren J, Iles MM, Sjölander A, Li Y, Chia KS, Liu E, Hall P, Liu J, Wedrén S.
Linkage Disequilibrium Mapping of _CHEK2_: Common Variation and Breast Cancer Risk.
PLoS Medicine 2006; 3(6):e168.

III. Einarsdóttir K, Rosenberg LU, Humphreys K, Bonnard C, Palmgren J, Li Y, Li Y, Chia KS, Liu ET, Hall P, Liu J, Wedrén S.
Comprehensive analysis of the _ATM, CHEK2_ and _ERBB2_ genes in relation to breast tumour characteristics and survival: a population-based case-control and follow-up study.
Breast Cancer Research 2006; 8(6):R67

IV. Einarsdóttir K, Humphreys K, Bonnard C, Li Y, Li Y, Chia KS, Liu ET, Hall P, Liu J, Wedrén S.
Effect of _ATM, CHEK2_ and _ERBB2_ tagSNPs and Haplotypes on Endometrial Cancer Risk.
Human Molecular Genetics 2006; Advance Access December 12

CONTENTS

ABBREVIATIONS

A-T	Ataxia-Telangiectasia
ATM	Ataxia-Telangiectasia Mutated
BMI	Body Mass Index
CHEK2	Checkpoint Kinase 2
CI	Confidence Interval
CYP17	Cytochrome P450c17
DASH	Dynamic Allele-Specific Hybridization
ddNTP	Dideoxynucleotide Triphosphate
DNA	Deoxyribonucleic Acid
ERBB2	V-erb-b2 Avian Erythroblastic Leukemia Viral Oncogene Homolog 2
ER	Estrogen Receptor
ESR1	Estrogen Receptor 1
HR	Hazard Ratio
HWE	Hardy-Weinberg Equilibrium
kb	Kilobase
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PLEM	Partition Ligation Expectation Maximization
PR	Progesterone Receptor
RFLP	Restriction Fragment Length Polymorphism
SNP	Single Nucleotide Polymorphism
tagSNP	Haplotype Tagging Single Nucleotide Polymorphism
TNM	Tumour, Nodes, Metastases

INTRODUCTION

Whilst our knowledge of the human body is extensive, at the same time we are still ignorant in many areas. The fact that we are often restricted to simplified experiments – whether the experiments are epidemiological observations or controlled laboratory situations – might be one of the reasons for this ignorance. We repeatedly explore an isolated feature, a cause and effect. However, the human body and its cells do not work in such a simple way. Cells are a complex system of DNA, proteins and fats. All have their necessary role and all make the picture complete. We most often cannot isolate one gene, one protein or one fatty acid and assume that the effect we observe is the only effect that exists.

In each of my studies I examine one gene at a time, looking at a single effect on cancer risk and survival. Even though most of the genes are covered comprehensively with respect to common variation, I still only explore one gene at a time. As has been widely stated, cancer is not a simple process. It is most likely not an effect of any single gene in each case. One possible reason why in my studies I have failed to find a major association – even though my hypotheses were reasonable – might be that I am simply not looking at the whole picture. However, as is often said in Icelandic: ”Margt smátt gerir eitt stórt” (many small pieces bring light to the whole picture). This is my contribution to the scientific world, small pieces of knowledge meant to shed a light on the mechanisms of breast and endometrial cancer aetiology and progression.

This project initiated in the early 90's with questionnaire-based case-control studies on breast and endometrial cancer, where extensive information on all participants was obtained. In the late 90's, a subset of the participants in the initial questionnaire-based studies were selected to be included in genetic studies. Blood and tissue samples were collected, DNA extracted, candidate genes selected, and polymorphisms analyzed. Development in the field was rapid, however, and at the end of this first phase of the genetic studies the investigators believed that a larger number of candidate genes – preferably whole pathways – needed to be studied with better coverage of each gene. Fortunately, enough blood and tissue samples had been collected from the participants to support another phase of the genetic studies. At this point, in February 2003, I began my thesis work on this project. DNA was extracted a second time and candidate gene selection began. In my thesis, I decided to investigate the *ESR1, EGF, ATM, CHEK2* and *ERBB2* genes in relation to breast cancer risk. I had explored the relationship of 300 candidate genes with estrogen, estrogen metabolism and estrogen downstream pathways and chose these five to be the main focus of my thesis. In the end however, my thesis included studies of the *CYP17, ATM, CHEK2* and *ERBB2* genes in relation to risk of breast and endometrial cancer as well as breast cancer prognosis. *CYP17* was originally selected and genotyped in the previous phase of the genetic studies, but was included in my thesis. Analyses of the *ESR1* and *EGF* genes will however have to wait their turn.

During the four years of my studies, development in the field of genetic association studies has again been rapid – or should we say continues to be rapid. We have now the capacity to perform genome-wide association studies. We therefore do not necessarily have to rely on existing biological knowledge for selecting candidate genes. The question remains however whether genome-wide association studies will in the end provide us with the complete picture. Perhaps in combination with well powered linkage studies, we will be able to reveal the genetic spectra involved in cancer aetiology. We could say that we are at the beginning of a new era. We have recognized our mistakes in the past and are eager to move forward. We have also realized that we need to share information and join forces. What these new efforts will bring, only future can tell. However, we are certainly moving in the right direction.

BACKGROUND

Human Genome Variation

The haploid human genome constitutes about 3.3 billion basepairs [1]. Around 3% of the genome consists of coding sequences [1] including 30,000-40,000 protein-coding genes [2]. The majority of the genome (99.9%) is identical between any two individuals, but variation does exist between two copies of the same chromosome [3]. New mutations can arise in individuals, either in somatic cells or in the germ-line. However, because mutation rates are low, the vast majority of allele differences within an individual are inherited rather than resulting from somatic mutations [1]. Most mutations arise as copying errors during DNA replication because DNA polymerases are error-prone, but DNA can also be damaged by exposure to natural ionizing radiation and reactive metabolites.

Large-scale chromosome abnormalities involve loss or gain of chromosomes or breakage and rejoining of chromatids. Smaller scale mutations can be grouped into the following mutation classes [1]:

- **Base substitutions** involve replacement of usually a single base.
  - **Splice site mutations** create or destroy signals for exon-intron splicing.
  - **Synonymous (silent) mutations** result in a new codon specifying the same amino acid.
  - **Non-synonymous mutations**
    - ◊ **Nonsense mutations** – a codon specifying an amino acid is replaced by a stop codon.
    - ◊ **Missense mutations –** the altered codon specifies a different amino acid.
- **Deletions** – one or more nucleotides are eliminated from a sequence.
- **Insertions** – one or more nucleotides are inserted into a sequence.
- **Framshift mutations** that can be produced by deletions, insertions or splicing errors.
- **Variable number tandem repeat polymorphisms** – caused by deletions/insertions in tandemly repetitive DNA.
  - **Microsatellites** – very short di-, tri- and tetranucleotide repeats.
  - **Minisatellites –** repeats of intermediate length.

Mutations drive evolution, but they can also be pathogenic and cause or increase the risk of diseases. Pathogenic mutations can occur in coding sequences of genes, in intragenic non-coding sequences (necessary for correct expression of the gene), or in regulatory sequences outside exons (promoter elements or distantly located regulatory elements) [1]. Mutations can be broadly categorized into gain-of-function and loss-of-function mutations [1]. Gain-of-function mutations usually cause dominant phenotypes, as the presence of a normal allele does not prevent the mutant allele from behaving abnormally. Loss-of-function mutations, on the other hand, most often produce recessive phenotypes, which means that normal function can be sustained with half the amount of the genetic product. For some gene products, 50% of the normal level is not sufficient for normal function, a phenomenon called haploinsufficiency. Also, sometimes a non-functional mutant protein interferes with the function of the normal allele in a heterozygous person, giving a dominant negative effect.

Linkage vs. Association

Linkage analysis was widely used in the early 1990s to locate genes and mutations involved in monogenic disorders. In linkage analysis, co-segregation of a disease and marker alleles is assessed among related individuals. If evidence for such co-segregation is found, one can infer the existence of a disease-causing locus near the marker locus. Because the focus of linkage mapping is on the small number of generations within a family, a limited number of recombination events have occurred and linkage can be detected over large genetic distances [4, 5]. Hence, only approximately 300 highly informative microsatellite markers evenly spaced across the human genome are needed in a genome-wide scan [6]. Microsatellites are suited for linkage studies since they typically have a large number of alleles (i.e. repeat lengths), making it easy to identify alleles co-segregating with a disease.

Linkage studies have been useful in detecting rare mutations in genes with high penetrance and strong effects that are involved in monogenic disorders or complex diseases [6-8]. However, they are not very powerful in detecting genes with low penetrance and small to moderate effects on a disease. Recent focus has therefore been on population-based genetic association studies, as they can more easily detect small effects of low penetrance common alleles. Association studies however require a much denser set of markers than linkage studies since the size of the genomic regions harbouring alleles that co-segregate with a disease in the population may be very small [4, 5]. This is due to the large number of recombinations over successive generations in the population. As single nucleotide polymorphisms (SNPs) are by far the most abundant variants in the human genome [1], they have become the markers of choice in genetic association studies.

Single Nucleotide Polymorphisms

A SNP is defined as a single base change that occurs with a frequency over 1% in the population [1]. It has been estimated that about 11 million SNPs with at least 1% minor allele frequency exist in the genome [9]. They are found throughout the genome, e.g. in exons, introns, intergenic regions, in promoters or enhancers. However, SNPs in non-coding sequences and synonymous SNPs in coding sequences are generally more common than non-synonymous SNPs. For example, a typical gene contains only one or two missense SNPs [10].

There are several reasons for the usefulness of SNPs in genetic association studies in addition to their abundance in the human genome:

1. As SNPs are single base substitutions, they can be rapidly and efficiently genotyped.

2. Groups of adjacent SNPs may exhibit patterns of correlations that can be used to enhance gene mapping.

3. Not only can SNPs be used as surrogate markers – like microsatellites – but some of them can also be tested directly as functional variants.

4. SNPs are less mutable than other types of variants [11], which minimizes the possibility that associations will be confounded by alleles having mutated to different forms between generations.

Direct vs. Indirect Association

In genetic association studies, the SNPs of interest can either be tested directly or indirectly [12]. Direct testing implies that variants known to have deleterious effect on the protein product are targeted. This type of genetic association study most resembles classical epidemiology case-control studies where the frequency of the exposure (the genetic variant in this case) is compared between cases and controls. This genetic association study type is also the easiest to analyze and the most powerful, but the difficulty is the identification of candidate polymorphisms. It is likely that many causal variants involved in the development of common, complex diseases will be non-coding (see Genetic Spectra of Common Diseases below). Such variants may affect gene regulation and expression or differential splicing, but our ability to predict such effects are limited. However, even though direct genetic association studies only have the potential to identify a part of the genetic causes of common diseases, a whole-genome approach to direct association is worthwhile [12, 13].

In indirect genetic association studies, a marker (called a haplotype tagging SNP) inherited together (in linkage disequilibrium) with the causal locus can act as a surrogate for the causal locus and can thus be assessed directly instead of the causal locus. The analysis of indirect association studies is however not as straightforward as that of direct association studies, and indirect studies are also less powerful than direct studies. Furthermore, to be able to exclude that a causal locus exists when a negative association is observed in the area under study, the coverage of the genotyped SNPs in an indirect association study needs to be carefully assessed in order to ensure that the majority of the variation in the area has been predicted [12].

The main advantage of indirect association studies over direct association studies is the detection of causal variants that have not previously been identified in the human genome. If indirect association studies are well designed and have enough SNP coverage of the area under study, they have the potential to detect any common variant in the area that is associated with the disease. The whole-genome approach to indirect association studies has therefore enormous promise in the future for identifying common alleles that play a role in the development of common, complex diseases. In spite of that, candidate gene studies will continue to play an important part since they will allow genotyping of markers more densely, thus improving detection of true causal association as well as increasing the confidence that negative findings represent true negatives [12].

Linkage Disequilibrium

Linkage disequilibrium (LD) implies that a particular allele at one locus is found together on the same chromosome with a specific allele at a second locus more often than expected if the loci were segregating independently in a population [14]. Pairs of loci in LD are generally close together, but the distances also vary (Figure 1) [15].

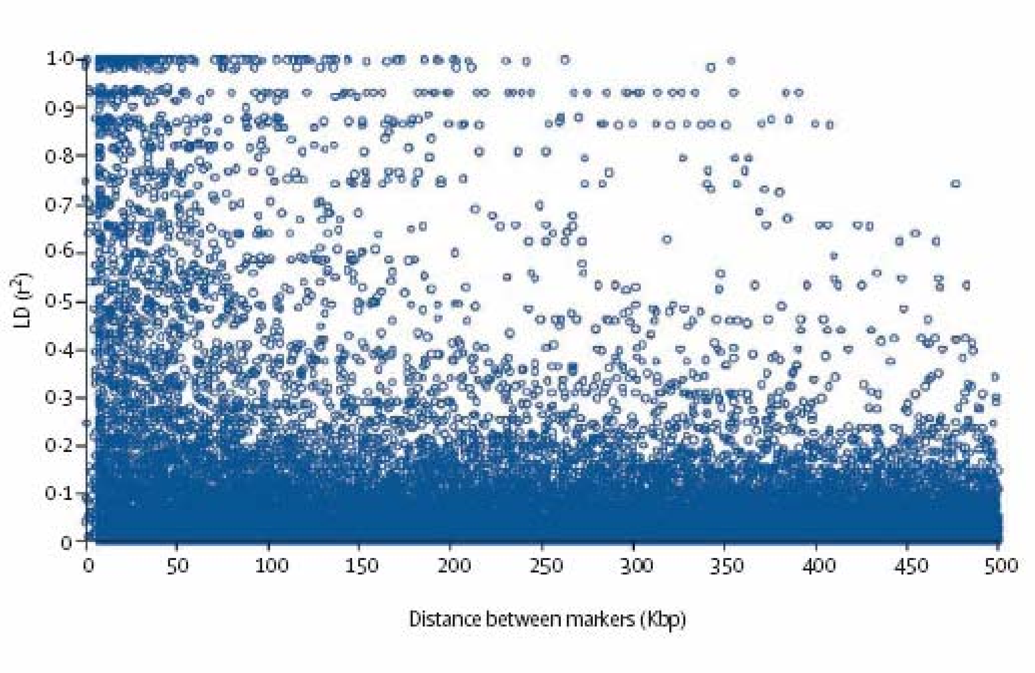

Figure 1. Observed patterns of LD decay on chromosome 22. (From Palmer et al. [15]).

When a variant is first formed in a population because of mutation, it will be perfectly correlated with nearby variants. However, over successive generations, recombination will break up the correlations and LD will decay. Mutation and recombination appear to have the most evident impact on LD, but demographic aspects of a population also contribute to the extent and distribution of LD [14].

The two most common measures of pairwise LD are the *D´* and R^2. Both measures range from 0 (no disequilibrium) to 1 (complete disequilibrium), but their interpretation differs. *D´* essentially measures the amount of recombination between loci, where *D´*=1 means that no recombination has occurred. Values of *D´* <1 indicate that LD has been disrupted by recombination, but values of *D´* between 0 and 1 have no clear interpretation [14]. Hence, LD based on *D´* can be divided into strong LD (*D´* near 1), weak LD (*D´* significantly lower than 1) and intermediate/unknown LD (intermediate *D´*) [16].

The R^2 measures the statistical correlation between two loci [16] and is related to the allele frequencies of the loci [12]. R^2 =1 means that knowledge of alleles at one locus can perfectly predict the alleles at the other [14]. Intermediate values of R^2 are easily interpretable since they are related to a) how well one locus predicts the other and b) the loss of power due to testing the locus of interest indirectly [14]. When using a surrogate marker for assessing an effect of a causal locus on disease, the sample size has to be increased by $1/R^2$ (where R^2 represents how well the surrogate marker predicts the causal locus) to achieve the same power as if the causal locus had been assessed directly [17]. These qualities of the R^2 measure make it the most appropriate measure of LD for the selection of haplotype tagging SNPs (tagSNPs) in genetic association studies.

The pattern of LD across the human genome has been described as a series of high LD regions (blocks) separated by short discrete segments of very low LD (recombination hot spots) [18, 19]. The high LD regions have been suggested to contain limited haplotypic diversity, where only three to five common haplotypes can account for 80-90% of all chromosomes in the population [19, 20]. However, discussion has surfaced regarding whether haplotype blocks have clear boundaries caused by recombination hot spots or whether they arise as a result of random recombination [14, 16, 21]. Furthermore, the idea of a clearly structured genome with respect to LD has been criticized and some have showed that patterns of LD can vary greatly [22].

Although the idea of tagSNPs was inspired by the observation of haplotype blocks – i.e. that few tagSNPs can predict the few common haplotypes observed in blocks – the tagging approach does not require clear blocks of LD [23, 24]. Regardless of how discreet the LD pattern is or what causes the pattern, the question important to genetic association studies is how well a tagSNP can represent the tagged SNP. Goldstein and colleagues argue that the average marker density necessary to find a set of tagSNPs sufficient to represent the common allelic variants in the human genome is 1-2 per 10kb [23]. This suggested marker density for the tagging approach in genetic association studies is somewhat lower than the marker density necessary to achieve stable block definitions, due to the fact that the length of haplotype blocks decreases with increasing marker density [16].

Genetic Spectra of Common Diseases

The current focus of genetic association studies exploring common diseases has almost entirely been on genetic markers that are common (minor allele frequency >1%). This is mainly due to two reasons. First, the 'common disease/common variant' (CD/CV) hypothesis [25]. This hypothesis states that genetic influences on diseases of high population prevalence are old, and are thus typically very common. This hypothesis is supported by a few examples, including the APOE ε4 allele in Alzheimer's disease [26], Factor V^{Leiden} in deep venous thrombosis [27], and PPARγ Pro12Ala in type II diabetes [28]. However, alternative hypotheses to the CD/CV exist, such as the classical disease heterogeneity hypothesis (or multiple rare-variant hypothesis), in which disease susceptibility is due to distinct genetic variants in different individuals and disease-susceptibility alleles have low population frequencies [29]. As common diseases are assumed to be influenced by many genetic and environmental factors, all with a modest effect on the trait (see below), genetic association studies are not amenable to discovering these rare alleles. The reasons are that the sample sizes required to detect the modest effects of the rare alleles will become impossibly large and since many patients will have unique mutations, associating each mutation with the disease will be almost impossible [15]. Hence, the second reason why the current focus of genetic association studies is on common genetic markers is a purely practical one. Genetic association studies can not possibly detect the modest effect of rare mutations on disease. Furthermore, since family-based linkage analyses detect mainly genes with strong effects on disease they will also be ill-equipped to locate rare alleles with small effects.

Wang and colleagues suggest that the allelic spectra of most common diseases probably falls between these two extreme hypotheses [30]. Empirical evidence suggests that both high- and low-frequency variants contribute to common diseases [26, 27, 31-34]. In addition, studies have indicated that the distribution of phenotypic-effect sizes of genetic variants is consistent with the existence of few genetic loci with large effects and numerous loci with small effects [30, 33]. The potential for a large number of variants with small effects to be involved in the aetiology of common diseases is supported by recent findings that allelic variation frequently affects gene expression and exon splicing [35-37]. This kind of variation is likely to have smaller effects than polymorphisms that affect the coding sequence. For example, causal alleles for monogenic disorders are highly penetrant and often cause severe changes in protein function. These mutations are often subject to negative selection and thus remain rare in the population. On the other hand, alleles that underlie complex diseases have more subtle effects on disease risk since they are more likely to include non-coding regulatory variants with a modest impact on expression. These alleles are therefore far less likely to be subject to strong negative selection and will thus most likely become more common in the population.

Breast Cancer

Cancer of the female breast is both a common and complex disease. It is responsible for one in ten of all new cancers diagnosed worldwide each year [38] and is the most common cancer in women in both high-resource and low-resource countries. It is also the most frequent cause of cancer death in females worldwide [38]. In Sweden, one in ten women will develop breast cancer by the age of 75 [39]. The age-standardized incidence rate was 142 per 100,000 women in 2004 with an annual increase of 1.6% for the previous 10 years [39]. In contrast to the increasing incidence, the breast cancer mortality has decreased slightly since 1975, which is thought to be due to intensified screening and improved therapy [39]. The 5-year survival rate has been increasing over the last several decades and is now estimated to be 85% [40].

The highest breast cancer incidence rates in the world occur in Northern and Western Europe, Northern America, Australia, New Zealand, Uruguay and Argentina [41]. Incidence is however low throughout Africa and Asia, as well as most of Central and South America [41]. Despite this, incidence rates have been rapidly increasing in some low risk areas, reflecting changes towards a ‘westernized’ lifestyle and reproductive pattern [42].

Factors that have been shown to increase breast cancer risk are late age at menopause [43], early age at menarche [43], hormonal replacement therapy [44, 45], recent oral contraceptive use [46], postmenopausal obesity [47], adult weight gain [48], and alcohol consumption [49], whilst premenopausal obesity [47] and physical activity [50, 51] reduce the risk. A common thread through these risk factors is that increased exposure to estrogens and progesterone seems to increase breast cancer risk whilst decreased exposure decreases the risk. Nulliparity and older age at first birth also increase the risk of breast cancer [43], whilst early age at first birth [43], long duration of breast feeding [52] and higher parity decrease the risk [43]. This is thought to be related to the protective effects of a fully differentiated mammary gland reached at full term pregnancy [53]. Other factors have been also suggested to increase breast cancer risk. Adolescent exposure to ionizing radiation is thought to affect breast cancer risk through DNA damage [54]. History of proliferative benign breast disease increases breast cancer risk and seems likely to be a precursor of the subsequent breast cancer if the cancer occurs within 10 years of the development of the benign breast lesion [55]. With a longer interval, however, benign breast disease seems to be merely a marker of increased susceptibility to breast cancer [55]. Diabetes type 2 seems to increase breast cancer risk through increased insulin levels that stimulate androgen synthesis and thereby cause decreased levels of sex hormone binding globulin and increased levels of free estrogen [56]. Tall stature appears to increase breast cancer risk which might be due to the fact that tall women may develop a higher number of ductal stem cells *in utero* than other women [47]. High mammographic breast density is also a risk factor for breast cancer [57]. Epithelium and stroma appear white on a mammogram, a phenomenon which is referred to as mammographic density [57]. High mammographic density has been associated with a greater total nuclear area, a greater proportion of collagen, and a greater area of glandular structures, which might reflect a greater number of breast cells at risk [57].

A positive family history of breast cancer is one of the major known risk factors for the disease. Women with one affected relative have an approximately twofold increased risk of breast cancer compared to women with no affected relatives and the risk increases with increasing number of affected first-degree relatives [58-60]. These observations suggest that heritable factors are important in breast cancer aetiology. Indeed, a Nordic twin study has reported that hereditary factors explain 27% of breast cancers [61].

BRCA1 and BRCA2

An intensive search for genetic factors causing hereditary breast cancer has been ongoing for the last couple of decades. Initially this search led to the identification of the *BRCA1* and *BRCA2* genes. *BRCA1* was localised to chromosome 17q21 by genetic linkage in 1990 [8] and subsequently cloned in 1994 [62]. *BRCA2* was localised to chromosome 13q12-13 in 1994 [63] and cloned in 1995 [64, 65]. Mutations in these genes are highly penetrant but the prevalence of specific mutations is small. Most mutations found in breast and/or ovarian cancer families appear to truncate the protein product [66]. The majority are small frameshift insertions or deletions, but missense and nonsense alterations as well as mutations affecting splice sites have also been described [66].

A number of population-specific founder mutations have been identified in the *BRCA* genes. In *BRCA1* the two most common mutations are 185delAG and 5382insC [67] which occur at a 10-fold higher frequency in the Ashkenazi Jewish population than in non-Jewish Caucasians [68]. Founder mutations in *BRCA2* include the 6174delT mutation – which has been found in Ashkenazi Jews [69] – and the 999del5 mutation, which has been found in the Icelandic and Finnish populations [70, 71]. The most frequent mutation in Sweden is the 3171ins5 in *BRCA1,* which has been reported in 65% of all breast cancer families detected with *BRCA1* or *BRCA2* mutations in Western Sweden [72]. In most countries a higher proportion of breast cancers seem to be due to *BRCA1* mutations than to *BRCA2* mutations [73]. This is in contrast to observations in Iceland where *BRCA2* linkage has been shown to predominate [71, 74].

BRCA1 and *BRCA2* are thought to account for a high proportion of high-risk breast cancer families. Ford and colleagues assessed the contribution of *BRCA1* and *BRCA2* to inherited breast cancer by linkage and mutation analysis in 237 families with four or more cases of breast cancer diagnosed below age 60 years [75]. They reported that 84% of the families were linked to either *BRCA1* or *BRCA2*. The majority (81%) of the breast-ovarian cancer families showed linkage to *BRCA1*, with most others (14%) showing linkage to *BRCA2*. On the other hand, the majority of families with male and female breast cancer showed linkage to *BRCA2* (76%). Furthermore, the estimated cumulative risk of breast cancer by the age of 70 reached 84% for *BRCA2* mutation carriers [75] and 71% for *BRCA1* carriers [76].

Estimates based on high-risk families such as reported by Ford and colleagues [75] are directly relevant to high-risk families, but may overestimate the risk in carriers randomly selected from the population. A recent meta-analysis reported penetrance estimates from 22 population studies [77]. The average cumulative risks of breast cancer by age 70 years were 65% in *BRCA1* mutation carriers and 45% for *BRCA2* mutation carriers.

As mentioned above, the prevalence of *BRCA1* and *BRCA2* mutations is 84% among high-risk breast cancer families [75]. However, the prevalence in families with fewer breast cancer cases is markedly lower; excess familial risk due to mutations in these genes has been estimated to be only 17% [34]. The most likely explanation is that whilst *BRCA1* and *BRCA2* are the most important high penetrance breast cancer susceptibility genes, a collection of other genes conferring lower risks explain the majority of the familial aggregation [33].

Five percent of breast cancer cases with first degree family history carry mutations in *BRCA1* and *BRCA2* [34]. Numerous studies have sought to estimate the overall frequencies of *BRCA1* and *BRCA2* mutations in breast cancer cases unselected for family history, but the majority of the studies have only included women at young age of diagnosis. The overall mutation prevalence in breast cancer cases under 55 years of age as been reported to be 0.7% for *BRCA1* and 1.3% for *BRCA2*, but was significantly higher in cases diagnosed under 35 years of age [34]. Thompson and Easton estimate the overall fraction of breast cancer cases in outbred populations carrying *BRCA1* and *BRCA2* mutations to be around 1-2% for each gene [66]. The overall frequencies of *BRCA1* and *BRCA2* mutations within large outbred populations have been inferred to be somewhat lower; 0.13% for *BRCA1* and 0.17% for *BRCA2* [78].

Endometrial Cancer

Uterine endometrial carcinoma is the 7th most common cancer in females worldwide and the 14th leading cause of cancer deaths [38]. Incidence rates are higher in North America and Europe than in Asia and Africa [79]. In Sweden, 3 in 100 women develop endometrial cancer by the age of 85 [39]. The age-standardized incidence has been slowly increasing for the last five decades and was 26 per 100,000 women in 2004 [39]. The mortality rate has on the other hand been decreasing slightly over the past four decades with the 5-year survival rate currently being 82% [80].

Endometrial cancers can be divided into Type I endometroid tumours and Type II non-endometroid tumours [81-83], where endometroid tumours constitute the majority of endometrial cancers. The endometroid tumours appear to be the tumours that are mainly caused by estrogen exposure [81-83] since they are associated with endometrial hyperplasia, express estrogen and progesterone receptors [84] and are associated with elevated levels of serum estradiol [82]. They are also characterized overall by a favourable prognosis [81-83].

The most important risk factor for endometrial cancer is unopposed estrogen exposure. This is reflected in the increased endometrial cancer risk in women with late age at menopause or adult obesity [85], in women that use postmenopausal estrogen without progestin addition [86] or in women that are nulliparous [87]. It is also reflected in the decrease in risk in women that use combined oral contraceptives [88], have higher parity [87], smoke [89] or are physically active [90]. Another factor that increases the risk is diabetes [85], whilst late age at last birth decreases the risk [87]. The underlying biological mechanisms for the last two associations are not clear but a few have been suggested. The biological mechanism in diabetics may be similar for breast and endometrial cancer, i.e. that increased insulin levels lead to increased levels of free estrogen [56]. For women delivering their last child late in life it has been suggested that the birth may provide protection by mechanically clearing the uterine lining from cells that have undergone malignant transformation [91].

Another important risk factor for endometrial cancer is family history of the disease. Having a first degree relative with endometrial cancer approximately doubles the risk of developing the disease [87]. This observation suggests that endometrial cancer is not only caused by environmental factors, but that genetic components also play a role in the disease development.

Genetics of Endometrial Cancer

Hereditary non-polyposis colorectal cancer (HNPCC) is a dominantly inherited syndrome with germ-line abnormalities in one of five DNA-mismatch repair genes (*MSH2, MLH1, PMS1, PMS2, MSH6*) with resultant microsatellite instability. Females with HNPCC have a tenfold increased lifetime risk of endometrial cancer compared with that of the general population and the lifetime risk of endometrial cancer (60%) is higher than that for colorectal carcinoma (54%) [92].

The lifetime risk of endometrial cancer in HNPCC families related to specific mutations in the DNA-mismatch repair genes has been explored in a few studies. Vasen and colleagues reported 35-40% cumulative risk of endometrial cancer by age 70 in women from 40 HNPCC families carrying *MSH2* mutations and a 25% cumulative risk in female *MLH1* mutation carriers from 34 HNPCC families [93]. Hendriks et al. examined 20 HNPCC families with mutations in *MSH6* and reported mutation carriers to have 71% cumulative risk of endometrial cancer by the age of 70 [94].

Somatic mutations are common in endometrial cancers but differ between Type I and Type II cancers. Type I carcinomas frequently show mutations in the *MLH1, MSH2, MSH6, PTEN, KRAS* and *ß-catenin* genes, whilst Type II cancers are more likely to contain amplification of the *ERBB2* gene and mutations in the *TP53* gene [95].

Candidate Genes

In this population-based, genetic association study, we applied both direct and indirect analyses to investigate the *CYP17, ATM, CHEK2* and *ERBB2* genes in relation to breast cancer risk, breast cancer survival and endometrial cancer risk. The four genes have previously been extensively studied in relation to breast cancer. However, since we did not study the effect of *CYP17* on endometrial cancer risk and since common variation in the *ATM, CHEK2* and *ERBB2* genes has never before been explored in relation to the risk of this cancer, I focus my literature review below on the genes' relationship with breast cancer risk as well as breast cancer survival when applicable.

CYP17

One of the key enzymes in the synthesis of sex hormones, such as estrogens and androgens, is Cytochrome P450c17 (CYP17). CYP17 catalyses the 17α-hydroxylation of pregnenolone and progesterone, and these intermediates are then converted to dehydroepiandosterone and androstenedione by the 17,20-lyase activity of the enzyme [96, 97].

The *CYP17* gene spans 7 kb of genomic sequence on chromosome 10 (dbSNP build 126). A single T (A1 allele) to C (A2 allele) base change in the 5' promoter region of *CYP17* (c.1-34T>C) has been suggested to create an additional binding site for the transcription factor Sp1 [98]. This could theoretically lead to increased levels of the enzyme, but use of the extra binding site has not been confirmed experimentally [99]. Three groups have reported an association of the A2 allele with increased breast cancer risk [100-102] whilst others have failed to replicate those findings [99, 103-115]. A recent meta-analysis of 15 case-control studies did not find any overall association [116] but this analysis was criticized by Feigelson and colleagues who found a borderline significant association between the *CYP17* polymorphism and advanced breast cancer [117].

ATM

The *ATM* (*ataxia-telangiectasia mutated*) gene covers 146.3 kb on chromosome 11 (dbSNP build 126). The ATM protein is activated in response to DNA damage and triggers phosphorylation of CHEK2 and other proteins that promote cell cycle arrest and activation of DNA repair [118-124]. The *ATM* gene is mutated in ataxia-telangiectasia (A-T), a rare autosomal recessive disorder associated with a complex phenotype that includes radiosensitivity and increased risk of cancer. Breast cancer risk has been found to be increased in relatives of A-T patients [125, 126] and in A-T heterozygotes [127, 128]. Previous mutation screening studies have indicated that missense mutations in *ATM* – not protein truncating mutations – are over-represented in breast cancer cases compared to the general population [129-134]. A recent publication refuted this finding and found that *ATM* mutations that cause ataxia-telangiectasia – i.e. truncating, splicing and missense mutations – are breast cancer susceptibility alleles [135]. They found an over twofold increase in breast cancer risk related to a combination of 12 mutations, six of which were truncating mutations [135]. Thus, controversy remains both regarding which type of mutations in *ATM* are involved in breast cancer aetiology, and also which mutations actually drive the association with breast cancer risk [134, 136-144]. With regard to common variation in *ATM*, three groups have reported a significant effect between specific common haplotypes in *ATM* and breast cancer risk [141, 144, 145], whilst one group did not find any association [146]. No association has been reported between *ATM* common haplotypes and breast cancer survival or tumour characteristics.

The rare 4258 C→T and 2527 T→C mutations in *ATM* are two of many missense variants that have been thought to increase breast cancer risk. Although the 4258 C→T and 2527 T→C missense variants do not appear to target residues known to be crucial for ATM function [134], increasing evidence suggests that missense variants in *ATM* cause chromosomal instability and abolish the radiation-induced kinase activity of ATM [147]. Mutant ATM protein also appears to interfere with normal ATM function in a dominant-negative manner [147].

CHEK2

The *CHEK2* (*checkpoint kinase 2*) gene spans 54.1 kb on chromosome 22 (dbSNP build 126). Following phosphorylation of the CHEK2 protein by ATM after DNA damage, activated CHEK2 phosphorylates p53, Cdc25, and BRCA1, thereby promoting cell cycle arrest and activation of DNA repair [118-121]. Mutations in the *CHEK2* gene have been found in patients with Li-Fraumeni syndrome [148], a highly penetrant familial phenotype characterized by the occurrence of breast cancers, sarcomas, brain tumours, leukemias, and adrenal cortical tumours [149]. One of the mutations – a rare deletion in *CHEK2* named *1100delC* – leads to a premature termination of translation that abolishes CHEK2 kinase activity [150]. This deletion has been found to increase breast cancer susceptibility at the population level [32] and in families without *BRCA1* or *BRCA2* mutations [151, 152]. It has also been associated with breast tumours of high grade [153, 154] as well as steroid receptor positive breast tumours, but not with overall survival [153]. Other mutations in the *CHEK2* gene have been inconsistently related to breast cancer risk [155-161]. Of the two common *CHEK2* polymorphisms studied so far, neither was associated with breast cancer risk [162]. Nothing has been reported regarding *CHEK2* common variation in relation to breast cancer survival or tumour characteristics.

ERBB2

The *ERBB2* (also named *HER2*) gene covers 33.7 kb on chromosome 17 (dbSNP build 126). The ERBB2 protein is a transmembrane glycoprotein with tyrosine kinase activity [163-167] that plays a major role in signal transduction pathways. Activation of the pathways results in a variety of cellular responses, including proliferation, cell differentiation, cell motility and survival [168-170]. The *ERBB2* gene has been shown to be often amplified and/or over-expressed in breast and endometrial cancers [171-176]. This hyper-activation of the receptor appears to be a marker of tumour aggressiveness and is associated with poor breast and endometrial cancer prognosis [171-173, 175-177]. Most previous publications regarding *ERBB2* and breast cancer risk have explored one common variant, I655V, but results have been inconsistent [178-181]. Two groups have studied common haplotypes in *ERBB2* in relation to breast cancer risk, but found no association [178, 181]. One group has investigated *ERBB2* common variation in relation to breast cancer survival and found an association between poor breast cancer prognosis and *ERBB2* common haplotypes [181].

Biological Hypotheses

The *CYP17* and *ERBB2* genes are obvious role players in both breast and endometrial cancer development. Estrogen exposure is an important risk factor for both cancers, so a gene involved in estrogen metabolism like *CYP17* is a clear candidate. Since the *ERBB2* gene is amplified and/or over-expressed in breast and endometrial tumours, a variant in the gene affecting this amplification could be involved in breast or endometrial aetiology or survival.

Mutations in the *ATM* and *CHEK2* genes appear to be involved in breast cancer development, so common variation in the genes could also play an important role. Furthermore, variation in the *ATM* and *CHEK2* genes could affect breast cancer survival through increased radiosensitivity [182-184]. But why study these genes specifically in relation to endometrial cancer in addition to breast cancer when they could be involved in the development of any cancer via their role in DNA repair? As mentioned above, estrogen exposure is the main risk factor for endometrial cancer. Estrogen metabolites have been reported to cause a number of DNA lesions [185], among which are double strand DNA breaks [186]. DNA double strand breaks appear to be the predominant signal for the activation of pathways mediated by the ATM protein [187]. Once activated, the ATM protein triggers phosphorylation of CHEK2 [118]. Defects in the *ATM* and *CHEK2* genes could thus be involved in endometrial cancer development via their role in DNA damage checkpoint regulation, especially in combination with increased estrogen exposure.

AIMS

General aims:

We wanted to assess whether common variation in selected candidate genes was involved in breast or endometrial cancer development, either as main effects, or by interacting with environmental cancer risk factors. We were also interested in whether carriers of common variants in the selected genes would be prone to develop breast tumours of certain characteristics or would have increased risk of breast cancer death.

Specific aims:

Paper I

1. To examine whether the *CYP17* c.1-34T>C polymorphism affects breast cancer risk overall, or in combination with environmental breast cancer risk factors.

Papers II-IV

1. To assess common variation in the *ATM, CHEK2* or *ERBB2* genes in relation to:
   - Overall breast cancer risk
   - Interaction with breast cancer risk factors
   - Breast cancer survival
   - Risk of tumour-characteristics defined breast cancer

   - Overall endometrial cancer risk
   - Interaction with endometrial cancer risk factors
2. To investigate whether rare variants in *ATM* and *CHEK2* affect overall breast or endometrial cancer risk.

SUBJECTS

Parent Studies

Two large case-control studies were initiated in the early 90's to examine the effect of menopausal hormone use on breast and endometrial cancer risk. The study base included all Swedish-born women between 50 and 74 years of age and resident in Sweden between October 1993 and March 1995 (breast cancer study) or between January 1994 and December 1995 (endometrial cancer study). During these periods, all breast and endometrial cancer cases were identified at diagnosis through the six regional cancer registries in Sweden. Controls were randomly selected from the Swedish Registry of Total Population to match the cases in 5-year age strata. Most of the controls were shared between the studies (n=2633), but additional controls (n=735) were sampled after completion of the breast cancer study to match the recruitment period of the endometrial cancer study.

Patients received a mailed questionnaire after having been asked to participate by their respective physicians. The questionnaire included detailed information on intake of menopausal hormones and oral contraceptives, weight, height, reproductive history, medical history and other lifestyle factors. The average interval from diagnosis to data collection was 4.3 months. Controls were contacted directly with the questionnaire. In the endometrial cancer study, only women with an intact uterus were eligible as controls. Histological specimens for the endometrial case women were retrieved from all 35 pathology departments in Sweden and reviewed and re-classified by the study pathologist.

Participation rates in the parent studies are given in Table 1. Reasons for non-participation among cases and controls were poor health of controls, physicians' refusal (because of psychiatric disorder, death, anxiety or poor physical health of the patients), subjects' refusal (either to being approached at all or to return the questionnaires) or failure in contacting the woman.

Table 1. Participation rates in the breast and endometrial cancer studies.

	Breast cancer study		Endometrial cancer study	
	Cases	Controls	Cases	Controls
Parent studies				
Eligible	3979	4188	1055	4216
Participated	3345 (84%)	3454 (82%)	802 (76%)	3550 (84%)
Present studies				
Eligible[a]	2818[a]	3111[a]	802[b]	3550[b]
Selected	1801	2057	802	2074
Participated	1596 (89%)[c]	1524 (74%)	719 (90%)	1574 (76%)
Available DNA	1590[c]	1518[c]	716	1567

[a] Pre-menopausal women and women with previous malignancies excluded
[b] Women with previous malignancies excluded
[c] Numbers not applicable to Paper I

Present Studies

From the parent breast cancer study, we randomly selected 1500 breast cancer cases and 1500 age-frequency matched controls among the postmenopausal participants without any previous malignancy (except carcinoma *in situ* of the cervix or non-melanoma skin cancer). Similarly, we selected all 802 endometrial cancer cases and randomly selected 802 age-frequency matched controls among the pre- or postmenopausal participants in the endometrial cancer study without any previous malignancy. Women with previous cancers were excluded in order to minimize the risk of including a metastasis from a previous cancer in our study instead of an incident cancer. Cervical carcinoma in situ and non-melanoma skin cancer should not be metastatic so women with these conditions were not excluded.

With the intention of increasing statistical power in subgroup analyses, we further selected all remaining breast cancer cases, breast cancer controls and endometrial cancer controls (191 cases, 108 controls and 277 controls, respectively) who had used menopausal hormones (estrogen alone or any combination of estrogen and progestin) for at least 4 years (breast cancer study) or at least 2 years (endometrial cancer study). We also included all remaining participants (110 breast cancer cases, 104 breast cancer controls and 124 endometrial cancer controls) with self-reported diabetes mellitus. Since a large proportion of the controls were shared between the studies, we were able to add an additional 345 controls to the breast cancer study that had been selected for the endometrial cancer study as well as add a further 871 controls to the endometrial cancer study that had been selected for the breast cancer study. Numbers for the total selected participants in the present studies are shown in Table 1.

We contacted all selected living women by mail. Those who gave informed consent received a blood sampling kit by mail. Whole blood samples were drawn at a primary health care facility close to the woman's home and sent to us by standard mail. The majority of the samples arrived at Karolinska Institute within 24 hours after phlebotomy. All blood samples were immediately stored at -20°C. For deceased cancer cases and those cases who declined to donate blood but consented to our use of tissue, we collected archival paraffin-embedded, non-cancerous tissue samples taken at breast cancer surgery (for example cancer-free lymph nodes, uterine tube or cancer-free myometrium). We acquired 70% (breast cancer study) and 65% (endometrial cancer study) of the requested tissue samples; the main reason for non-participation was unwillingness or lack of time at the respective pathology department to provide the tissue blocks. In total, we obtained blood samples and archived tissue samples for 1321 and 275 (247 included in Paper I) breast cancer patients, respectively, and 603 and 116 endometrial cancer patients (Table 1). We also obtained blood samples for 1524 breast cancer controls and 1574 endometrial cancer controls (Table 1). Mean time from diagnosis of cases to arrival of the blood and tissue samples at our department was 5 years. Reasons for non-participation included lack of interest in research, a negative attitude towards genetic research, old age, and severe disease or death. Population-based participation rates (taking into account the proportion that did not participate in the parent questionnaire study) were 75% and 61% for the breast cancer cases and controls respectively, and 68% and 64% for the endometrial cancer cases and controls.

Categorization of Questionnaire Information

We defined age at menopause as the age of the last menstrual period or age at bilateral oophorectomy, if this took place at least one year before data collection. If the last menstrual period was less than one year before data collection, the woman was considered premenopausal. Women who had been hysterectomized in the breast cancer study or who were menstruating due to hormone treatment were considered postmenopausal when they reached the age at which 90% of the study participants had reached natural menopause. Women that we classified as postmenopausal in this way were assigned an age at menopause corresponding to the mean age at menopause in our data according to case-control and current smoking status.

Conjugated estrogens, estradiol and other synthetic estrogens were classified as medium potency estrogens. Estriol was classified as low potency estrogen. We only included information on orally administered menopausal hormones in the present study.

Age was divided to 5-year age groups and adjusted for in all analyses. Breast and/or endometrial cancer risk factors that we analyzed for gene-environment interaction were: Age at menarche (≤12 years, >12–14 years, >14 years); age at menopause (<49 years, 49–52 years, >52 years); medium potency estrogen only use (never, <4 years, ≥4 years; or never, <2 years, ≥2 years); medium potency estrogen and progestin use (never, <4 years, ≥4 years; or never, <2 years, ≥2 years); estrogen and progestin cyclically (less than 16 days of progestins per cycle, most commonly 10 days) (never, <2 years, ≥2 years); estrogen and progestin continuously (19 or more days of progestins per cycle, most commonly 28 days) (never, <2 years, ≥2 years); low potency estrogen use (ever, never); age at first birth (≤24 years, 25–29 years, ≥30 years); age at last birth (≤26 years, 27-33 years, ≥34 years); parity (nulliparous, 1 child, 2 children, >2 children); body mass index one year prior to diagnosis (weight in kg/(height in meters)^2) (<25, 25-<28, ≥28); regular smoking for at least 1 year or ever having smoked over 100 cigarettes (yes, no); first degree family history in at least one relative (yes, no); use of combined oral contraceptives where estrogens and progestins were given concurrently in a monthly cycle (ever, never); and self-reported diabetes mellitus (yes, no).

Age at menarche and menopause were categorized into quartiles according to the distribution in the controls. The 1st category included the 1st quartile of the data, the 2nd category contained the 2nd and 3rd quartiles, and the 3rd category contained the 4th quartile. Menopausal hormone use, age at first and last birth, parity, and body mass index where categorized according to cut-offs that in previous studies have been shown to be informative with regard to the variables' influence on breast and/or endometrial cancer risk.

Breast Tumour Characteristics and Follow-up

We retrieved information on date and cause of death until December 31st 2003 from the Swedish Causes of Death Registry and on date of emigration from the Swedish National Population Registry. Follow-up time began at date of diagnosis and ended on December 31st 2003, or at date of death or emigration, whichever came first. From medical records, we collected information on tumour characteristics such as tumour size, lymph node involvement, grade (tumour differentiation), histological type and date of first distant metastasis. We obtained information on tumour estrogen and progesterone receptor content and S-phase fraction (i.e. the proportion of tumour cells in the DNA synthesis phase of the cell cycle) from seven laboratories around Sweden that routinely perform these tumour measurements for all of Sweden. At the time of the study, all seven laboratories used an enzyme immunoassay (Abbott Laboratories) on cytosol samples for analyzing estrogen and progesterone receptor content. This method was estrogen receptor alpha specific [188]. The laboratories reported either quantitative measurements (i.e. fmol receptor per µg DNA or mg protein, and percentage of cells in S-phase) or categorical (i.e. strongly positive, positive, weakly postitive or negative for receptor status and high, intermediate or low for S-phase). A rather high proportion of this information was missing, due to the fact that these measurements were not routinely performed in the mid 1990s.

We classified the tumour characteristics as follows: TNM stage: (1) Tumour size ≤20 mm and no regional lymph node metastases; (2) tumour size ≤20 mm and lymph node metastases, or tumour size 20-≤50 mm, or tumour size >50 mm and no lymph node metastases; (3) inflammatory breast tumour, or tumour size >50 mm and lymph node metastases; (4) distant metastasis within 90 days after diagnosis. Lymph node involvement: (Yes) At least one metastasized lymph node; (No) no metastasized lymph node. Grade: (1) High differentiation; (2) intermediate differentiation; (3) low differentiation. Estrogen and progesterone receptor status: (Positive) ≥0.05 fmol/µg DNA or ≥10 fmol/mg protein, or categorically strongly positive, weakly positive or positive; (Negative) <0.05 fmol/µg DNA or <10 fmol/mg protein, or categorically negative. S-phase fraction: (High) ≥9% or categorically high; (Low) <9% or categorically low. We combined TNM stage 3 with TNM stage 4 in all association analyses due to small numbers.

Endometrial Tumour Characteristics

Endometrial cancers can be divided into Type I endometroid tumours and Type II non-endometroid tumours [81-83], where endometroid tumours constitute the majority of endometrial cancers. Endometroid tumours can be further divided according to cell differentiation (grade). We defined grade as follows: Grade I tumours were defined as well differentiated carcinomas, with maximum 5% solid areas; grade II tumours as moderately differentiated, with 6-50% solid areas; and grade III tumours as poorly differentiated or entirely undifferentiated, with more than 50% solid areas. Myometrial invasion was classified as: (No) None or less than 50% of the myometrial thickness; (Yes) at least 50% of the myometrial thickness or through the serosa.

METHODS

Paper I

DNA Isolation

We isolated DNA from 3 ml of whole blood using a Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA) according to the manufacturer's instructions. From non-malignant cells in paraffin-embedded tissue, we extracted DNA using a standard phenol/chloroform/isoamyl alcohol protocol [189].

Genetic Analyses

Colleagues at the Department of Medical Sciences, Uppsala University and at the Unit of Molecular Toxicology at the Department of Environmental Medicine, Karolinska Institute carried out all genotyping. They used two methods for genotyping the c.1-34T>C (rs743572) variant in *CYP17*: Multiplex fluorescent solid-phase minisequencing [190]; and dynamic allele specific hybridization (DASH) [191]. Results from the two methods were validated with PCR-RFLP [192]. Twenty-four percent of the samples were analysed with both minisequencing and DASH, and the genotypes obtained were identical. Exact PCR conditions, primer sequences and allele detection methods can be found in the supplement to Paper I.

Minisequencing

The minisequencing method is based on the extension reaction where a primer is attached adjacent to the variant and one fluorescent labeled ddNTP, which stops the elongation, is added to the reaction. If the SNP is for example G/A then ddCTP and ddTTP are added, each with different fluorescence dye. The fluorescence signal will indicate what alleles are present in the sample. The minisequencing method could not be easily multiplexed (see 'Papers II-IV', 'Genetic Analyses') so the genotyping personnel switched to DASH.

DASH

In the DASH method a probe specific to the wild-type allele is hybridized to the DNA strand surrounding the SNP. If the mutant allele is present in the DNA sample, a mis-match will occur at the SNP site. Following addition of a double strand-specific dye, the sample is steadily heated and fluorescence signal is continually monitored. Rapid fall in fluorescence indicates the denaturing temperature of the double stranded DNA. The mis-match double stranded DNA will denature at a lower temperature than the wild-type double stranded DNA, which indicates that a mutant allele is present in the sample.

Statistical Analyses

Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated using conditional logistic regression models with the A1/A1 genotype of the c.1-34T>C variant as the reference group. We conditioned the models on age (due to matching) and on duration of medium potency estrogen only use, estrogen+progestin use and self-reported diabetes mellitus (due to our over-sampling scheme). We chose to condition because in the presence of association between any of the selection variables and *CYP17* genotype, estimates of the main effect of *CYP17* could be biased. We did not expect confounding by other known breast cancer risk factors as they most likely would be intermediates between the genetic variant and breast cancer. We nevertheless assessed whether c.1-34T>C was associated with any of the known breast cancer risk factors. The likelihood ratio test and the Wald test statistic were used to test for interaction.

Papers II-IV

Overview

We downloaded all reported SNPs in the *ATM*, *CHEK2* and *ERBB2* genes from publicly available databases and selected thereof the SNPs that were eligible for our study. The selected SNPs were genotyped in 92 randomly selected controls. From that genotype information we were able to estimate the LD across the genes, reconstruct haplotypes and select the tagSNPs that could predict the single locus and haplotypic variation in the genes. The tagSNPs were genotyped in the full set of cases and controls so as to evaluate the association of common tagSNPs or their haplotypes with breast or endometrial cancer risk or survival. We additionally genotyped rare variants in *ATM* and *CHEK2* in all cases and controls and evaluated their relation with breast or endometrial cancer risk.

DNA Isolation

The Swegene laboratories in Malmö (Sweden) extracted DNA from 4 ml of whole blood using the QIAamp DNA Blood Maxi Kit (Qiagen) according to the manufacturer's instructions. The extraction was complete in March 2004 and the DNA samples were shipped to the Genome Institute of Singapore for genotyping. DNA was extracted from non-malignant cells in paraffin-embedded tissue using a standard phenol/chloroform/isoamyl alcohol protocol [189] in June 2004 at the Genome Institute of Singapore. Numbers for successfully extracted DNA samples are shown in Table 1.

SNP Selection

At initiation of this study, SNP data from the International Hapmap project was still sparse, so we decided to select SNPs from publicly available databases and characterize LD as well as choose tagSNPs using our own study population.

In October 2003, we downloaded from Ensembl (http://www.ensembl.org/) all the available SNPs in *ATM*, *CHEK2* and *ERBB2* and their 5kb flanking sequences. From them we selected all validated SNPs with a minor allele frequency of at least 1%, aiming for a marker spacing of less than 5kb. At that time, Ensembl had three criteria for a SNP to be considered validated: a) By frequency (a SNP has genotype frequency data); b) by cluster (SNP has been submitted at least by two submitters); or c) by 2hit-2allele (every allele of the variant has been observed in at least two chromosomes (i.e. in two different samples of DNA)). To fill in the gaps that exceeded 5kb, we included non-validated SNPs.

Due to inadequate coverage in the genes after this first batch of SNP selection (see ‘Statistical Analyses’, ‘Coverage’ below) we downloaded and selected a second batch of

SNPs in the *ATM*, *CHEK2* and *ERBB2* genes – this time including their 10kb flanking sequences – in January 2005. This selection was done using an integrated database of the Genome Institute of Singapore (GISSNP), which contained SNP information from dbSNP (build 123, http://www.ncbi.nlm.nih.gov/SNP/) and Celera. In June 2005, a third batch of SNPs was subsequently selected in the *ATM* and *ERBB2* genes. This time around the GISSNP database included information from dbSNP build 124.

Reasons for incomplete coverage despite selecting a large number of SNPs for initial genotyping were: a) Primers could not be designed around all selected SNPs due to repetitive sequence surrounding the locus; b) not all SNPs were successfully genotyped in at least 85% of the samples due to problems with multiplexing (see 'Genetic Analyses' below); and c) not all successfully genotyped SNPs were polymorphic. For example, in *CHEK2* we selected 151 SNPs for genotyping in total, we were able to design primers for 55 SNPs (36%), 34 SNPs (23%) were successfully genotyped, and 23 SNPs (15%) were polymorphic.

We additionally selected for genotyping a number of non-synonymous SNPs and SNPs from conserved sequences across human, rat and mouse. Among those which were successfully genotyped in the 92 controls, most were very rare variants and were thus not genotyped in the full set of cases and controls.

Genetic Analyses

Genotyping was performed at the Genome Institute of Singapore. Thousands of SNPs in different genes were genotyped concurrently so genotyping was not limited to the SNPs in the *ATM, CHEK2* and *ERBB2* genes. We used the Sequenom primer extension-based assay (San Diego, California) and the BeadArray system from Illumina (San Diego, California) for genotyping. *CHEK2* was entirely genotyped with Sequenom, but *ATM* and *ERBB2* were also genotyped with Illumina.

All genotyping results were generated and checked by laboratory staff unaware of case-control status. All DNA plates (96 wells) contained two negative controls and two distinct positive controls. For genotypes to be considered accurate, 97% of each set of positive controls were required to show an identical genotype and 97% of the negative controls needed to be without contamination. Only SNPs where at least 85% of the samples gave a genotype call were analysed further. As quality control, we genotyped 200 randomly selected SNPs in the 92 control samples using both the Sequenom system and the BeadArray system. The genotype concordance was >99.5%, suggesting high genotyping accuracy.

Three major steps characterize most genotyping methods: 1) Initial amplification of the genomic template, 2) generation of allele specific products, and 3) detection of allele-specific products. High-throughput genotyping systems are created by combining a certain assay for generation of allele-specific products with a specific detection platform. The main strategies for creating high-throughput genotyping platforms are to analyze multiple samples at the same time (i.e. 96 well plates with one sample in each well) and to analyze multiple SNPs in a single well (called multiplexing).

Sequenom

Sequenom uses the primer extension assay for generating allele-specific products. In the extension reaction, the primers anneal adjacent to the SNP. Adding the correct mix of bases will make the primer adjacent to one allele of the variant elongate for one base. In heterozygous samples, the primer adjacent to the other allele elongates for two bases. This results in PCR products of different length depending of the allele present. The products are inserted into a mass spectrometer, which measures the time of flight of the products. The products containing only one extra base will fly faster and produce a peak in the output that is differentially positioned to the peak for the product with the two additional bases. We multiplexed our analyses by analyzing the SNPs in 1-plex assays up to 12-plex assays. With maximum throughput, the Sequenom platform can generate 18,400 genotypes per day.

Illumina

For higher throughput, we used the Illumina platform, which can generate 300,000 genotypes per day. The Illumina company designs plates that carry 96 array bundles, one for each sample of DNA. Each array bundle contains 46,080 beads that are divided into 1536 bead types with one bead type per SNP. Each bead type includes 30 beads that together carry oligonucleotides specific for one SNP. Hence, 1536 SNPs can be genotyped simultaneously for each DNA sample. Illumina uses the oligo ligation assay for generating allele-specific products. Three oligos are designed for each SNP locus. One oligo is specific to each allele of the SNP site. The third oligo hybridizes several bases downstream from the SNP site and carries a unique address sequence that targets the correct bead type for that SNP. The assay oligos hybridize to the genomic DNA sample. Extension of the appropriate allele-specific oligo and ligation of the extended product to the oligo with the unique address joins information about the allele present at the SNP site to the address sequence for the bead type. These DNA products are then dye-labeled depending on the allele present and hybridized to their complement bead type through the unique address. Different alleles of each SNP will give different fluorescence signals, which can be read from the bead types for each SNP with a 2-D reader. In each array bundle, the 30 beads for each SNP will give the same fluorescence signal as one array bundle corresponds to one DNA sample.

CHEK2 Duplicated Regions

Non-expressed duplications of the *CHEK2* 3' terminal exons and introns have been revealed on chromosomes 2, 7, 10, 13, 15, 16, X and Y [193]. We blasted our primers and probes for the *CHEK2* SNPs to the genome and found they could hit unique sequences on chromosome 22. However, with lesser scores, they could also hit the duplicated sequences on the other chromosomes. We observed that the SNPs on the duplicated regions tended to deviate from Hardy-Weinberg equilibrium (HWE) (four SNPs out of six). We therefore relied on HWE tests to identify the SNPs in *CHEK2* that seemed to be genotyped from more than one genomic location.

The *1100delC* was designed in monoplex on Sequenom with primers located outside the duplicated regions to ensure genotyping of the expressed copy of *CHEK2*.

Statistical Analyses

LD Characterization, Haplotype Reconstruction and tagSNP Selection

Using the Haploview program [194], we produced LD plots of the D′ values for the SNPs in *ATM, CHEK2* and *ERBB2* that were genotyped in the 92 controls. We reconstructed haplotypes for all three genes using the PLEM algorithm [195] implemented in the *tagSNPs* program [196] and selected tagSNPs based on the R^2 coefficient, which quantifies how well the tagSNP haplotypes predict SNP genotypes or the number of copies of haplotypes an individual carries. We chose tagSNPs so that common SNP genotypes (minor allele frequency ≥0.03) and common haplotypes (frequency ≥0.03) were predicted with $R^2 \geq 0.8$ [19].

Coverage

In order to evaluate our tagSNPs' performance in capturing unobserved SNPs within the genes and to assess whether we needed a denser set of markers to be genotyped in the 92 controls, we performed a SNP-dropping analysis [197, 198]. 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. A slightly more stringent value of 0.85 was used here, as we were predicting only SNPs and not haplotypes with our tagSNPs. We then estimated how well the tagSNP haplotypes of the remaining SNPs predicted the dropped SNP, which act as surrogates for unobserved SNPs in the gene. This evaluation can provide an unbiased and accurate estimate of tagSNP performance [197, 198].

HWE

The Hardy-Weinberg law predicts how gene frequencies will be transmitted from generation to generation given a specific set of assumptions. Specifically, if an infinitely large, random mating population is free from outside evolutionary forces (i.e. mutation, migration and natural selection), then the gene frequencies will not change over time and the frequencies in the next generation will be p^2 for the AA genotype, 2pq for the Aa genotype and q^2 for the aa genotype.

Shoemaker et al. stated that a population will never confirm exactly with the Hardy-Weinberg law [199]. However, in well designed genetic association studies, the conditions of HWE are generally applicable to the controls since (1) mating takes place at random with respect to genotype, (2) allelic frequencies are the same in males and females, and (3) mutation, selection, and migration are negligible [200]. Furthermore, deviation from HWE can lead to either false positive or false negative findings for association [200]. Errors in genotyping can easily lead to large deviations from HWE. With Sequenom for example, this can happen when DNA is of poor quality or assays perform poorly. Larger DNA products will then fly less efficiently in the mass spectrometer and will thus have lower intensity peak compared to the smaller products. This can cause homozygous excess, as heterozygotes are called incorrectly as the homozygous low mass allele, i.e. the two peaks for heterozygotes can not be distinguished. We manually checked our genotyping results for SNPs that deviated from HWE in the controls, and if the results could not be repaired, we excluded that particular SNP from all analyses.

We checked the assumption of HWE among the controls with the standard χ^2 test statistics using the observed genotype frequencies obtained from the data and the expected genotype frequencies obtained with the HWE. SNPs were regarded as deviating from HWE if $P < 0.01$. We relaxed the usual $P < 0.05$ cut-off in order to not exclude SNPs that deviated from HWE due to chance, which is frequently the case for SNPs with low minor allele frequency in samples of limited size.

Multiple Testing Correction

Our testing strategy was to fit a single model and to assess within each stratum of risk factor subgroup and for different tumour characteristics, haplotype-trait association as a global likelihood ratio test [201]. We accounted for the number of tests by using a permutation approach that controls the family wise error rate (probability of rejecting one or more true null hypotheses) and takes into account the dependence structure of the hypotheses [202]. Only when a haplotype global test was significant did we scrutinize the haplotype-specific effects.

Haplotype Dosage

We computed expected haplotype dosage using the *tagSNPs* program [196]. Haplotype dosages give estimates as to how many copies of a certain haplotype an individual carries. Assuming we have genotype information on two loci, an individual with two homozygous genotypes (e.g. AA and BB) will carry two copies of the same haplotype (AB). If the individual carries one heterozygous genotype (Bb) the individual will carry one copy of the haplotype AB and one copy of the haplotype Ab. However, if both loci are heterozygous, four possible haplotypes exist (AB, Ab, aB, ab) and the haplotype dosages will thus depend on the haplotype frequencies.

We computed the haplotype dosages with haplotype frequencies estimated for cases and controls combined, assuming Hardy-Weinberg equilibrium (HWE) of haplotypes. We then included the haplotype dosages as explanatory variables in our regression models. We assumed co-dominance of the haplotype effects in our analyses, i.e. the computed point estimates showed the risk increase associated with carrying one copy of a haplotype. The effect estimates should be squared in order to calculate the risk associated with carrying two copies of a haplotype.

Power

To estimate power in the risk analyses, we used a method for indirect genetic association studies described by Chapman et al. [203], which assumes co-dominant effects at an unobserved locus. To calculate power for log-additive effects in the survival analyses, we used the Quanto program [204] in a similar manner as Manolio et al. [205], i.e. by assuming two controls for each case.

Association Analyses

In Paper II, we applied conditional logistic regression models conditioned on age (in 5-year age-groups) as well as the selection variables (menopausal hormone use and diabetes) to assess the association of *CHEK2* tagSNPs or haplotypes with breast cancer risk. In Papers III and IV, we applied unconditional logistic regression models adjusted for age to assess the effect of *ATM, CHEK2* and *ERBB2* tagSNPs or haplotypes on risk of breast or endometrial cancer. Conditioning on the selection variables did not affect our estimates in Papers III and IV. The appropriateness of these approaches is argued for by Stram et.al. [196]. That is, when R^2 values are high, as is the case here, point and interval estimates obtained by this approach will be approximately accurate. We estimated the hazard ratio of death due to breast cancer in relation to the *ATM, CHEK2* and *ERBB2* tagSNPs or haplotypes using Cox proportional hazards models. To assess the proportional hazards assumptions of the Cox models we examined scaled Schoenfeld residuals and found no evidence against proportionality.

Confounding has been defined as the presence of a common cause to the exposure and the outcome [206]. We believe that lifestyle and reproductive breast cancer risk factors are unlikely to cause genetic variation in the genes, but they could be intermediates in the causal pathway between the genes and a) overall cancer and b) tumour characteristic-defined cancer. For completeness, we assessed among the randomly selected controls – using Kruskal-Wallis and Chi square tests – whether the tagSNPs where associated with known cancer risk factors.

RESULTS

Characteristics of Participants

The selected characteristics of the cases and controls participating in the present breast and endometrial cancer genetic studies reflected established associations (Table 2).

Breast cancer cases cases who participated via tissue sample donation were on average 1.5 years older (*P* = 0.0003) and were more likely to have been diagnosed with TNM stage 2 or more advanced cancers (*P* < 0.0001), compared to breast cancer cases who donated a blood sample. Endometrial cancer cases who participated via tissue sample donation were however 2.1 years older on average than endometrial cancer cases who participated by donating a blood sample (*P* = 0.002) and were more likely, though not significantly, to have poorly differentiated (grade 3) tumours (*P* = 0.08). Importantly, no significant differences in genotype frequencies were evident between those who participated via blood or tissue among the breast or endometrial cancer cases.

Table 2. Selected characteristics of the cases and controls participating in the present breast and endometrial cancer studies.

	Breast cancer study			Endometrial cancer study		
Characteristic	Number of cases/ controls	Cases/ Controls	*P*[a]	Number of cases/ Controls	Cases/ Controls	*P*[a]
		Mean			Mean	
Age (years)	1579/1516	63.3/63.1	0.405	705/1565	64.0/62.8	<0.0001
Age at menopause (years)	1569/1503	50.4/50.0	0.015	617/1506	51.0/50.1	<0.0001
Recent BMI (kg/m^2)[b]	1570/1495	25.8/25.5	0.073	704/1548	27.4/25.5	<0.0001
Age at first birth (years)	1341/1368	25.4/24.7	0.001	604/1406	24.6/24.7	0.634
Parity	1579/1516	1.8/2.2	<0.0001	705/1565	1.9/2.1	<0.0001
Duration of menopausal hormone use (years)		Percent			Percent	
0	1050/1085	67.1/72.7	---	498/1113	71.9/72.1	---
<4 (breast), <2 (endo)	206/190	13.2/12.7	---	49/131	7.1/8.5	---
≥4 (breast), ≥2 (endo)	308[c]/217[c]	19.7[c]/14.5[c]	---	146[c]/300[c]	21.1[c]/19.4[c]	---
Self reported diabetes mellitus (yes/no)	1577/1401	9.0[c]/7.8[c]	---	705/1443	10.1[c]/8.3[c]	---
Smoking (yes/no)[d]	1579/1516	42.8/42.7	0.998	705/1565	35.6/43.1	0.0008
Family history (yes/no)[e]	1540/1379	16.0/9.3	<0.0001	669/1399	10.2/5.1	<0.0001

[a] Kruskal-Wallis or Chi-square tests.
[b] One year prior to diagnosis.
[c] Long term users of menopausal hormones and women with diabetes mellitus were over-sampled.
[d] Regular smoking for at least 1 year or having ever smoked 100 cigarettes.
[e] Family history is defined as having at least one first degree relative with breast or endometrial cancer.

Paper I

We obtained *CYP17* c.1-34T>C genotypes from 1,544 breast cancer cases and 1,502 controls. The genotype frequencies were similar to previously published frequencies in Caucasian populations [99, 103-105, 114, 115] and did not deviate from Hardy-Weinberg Equilibrium (P = 0.927) among the controls. The *CYP17* c.1-34T>C was not associated with the environmental breast cancer risk factors among the randomly selected controls.

We found no overall association between *CYP17* c.1-34T>C and breast cancer risk, regardless of histopathology (Table 3). This negative result was not modified by menopausal hormone use or diabetes mellitus. We also did not detect any association when we considered stage 1 (n = 389) and stage 2 or more advanced (n = 591) breast cancers separately. None of these findings were altered by restricting the sample set to the randomly selected cases and controls or by including other breast cancer risk factors as co-variates in the logistic regression models.

In exploratory analyses, the A2/A2 genotype, compared to A1/A1, was associated with an increased breast cancer risk in women with age at menarche of 12 years or younger (P for interaction = 0.026; Table 3). Furthermore, carriers of the A2 allele conferred a decreased risk in women with menopause before 49 years of age compared to A1/A1 carriers (P for interaction = 0.062; Table 3). There was, however, no dose-response pattern in these findings and we therefore regarded them as being due to chance. Age at first birth, parity or body mass index did not seem to modify the non-association between *CYP17* c.1-34T>C genotype and breast cancer risk. All estimates remained unaffected after we restricted the analyses to the randomly selected cases and controls.

Table 3. *CYP17* c.1-34T>C in relation to breast cancer risk overall, restricted to histological type or stratified by breast cancer risk factors.

	CYP17 genotype					
	A1/A1		A1/A2		A2/A2	
	Cases/ controls[a]	OR[b] (CI)	Cases/ controls[a]	OR[b] (CI)	Cases/ controls[a]	OR[b] (CI)
All cancers	550/488	1 (reference)	711/638	1.0 (0.9-1.2)	238/212	1.0 (0.8-1.3)
Ductal cancers	420/488	1 (reference)	510/638	1.0 (0.8-1.1)	180/212	1.0 (0.8-1.3)
Lobular cancers	56/488	1 (reference)	90/638	1.3 (0.9-1.8)	24/212	1.1 (0.6-1.8)
Age at menarche (years)						
≤12	102/96	1 (reference)	138/128	1.1 (0.7-1.5)	74/38	1.9 (1.1-3.0)
>12-14	283/251	1.1 (0.8-1.5)	360/325	1.1 (0.8-1.5)	106/126	0.8 (0.6-1.0)
>14	107/94	1.1 (0.7-1.6)	152/126	1.2 (0.8-1.7)	41/38	1.1 (0.6-1.8)
Age at menopause (years)						
<49	145/123	1 (reference)	141/165	0.7 (0.5-1.0)	44/59	0.7 (0.4-1.1)
49-52	282/243	1.0 (0.7-1.3)	380/316	1.0 (0.8-1.4)	140/97	1.2 (0.9-1.8)
>52	119/117	1.0 (0.6-1.3)	184/150	1.1 (0.8-1.6)	54/56	0.8 (0.5-1.3)

[a] Including only cases and controls with complete information on menopausal hormones and diabetes mellitus.

[b] Analyses were conditioned on age, menopausal estrogen only use, use of estrogen in combination with progestin and diabetes mellitus.

Papers II-IV

Genotyping, LD Pattern and Coverage

Summary statistics on genotyping results and SNP coverage in the *ATM, CHEK2* and *ERBB2* genes are shown in Table 4. The SNPs successfully genotyped in 92 randomly selected controls are listed in Paper II (*CHEK2*) and in supplements to Papers III and IV (*ATM* and *ERBB2*).

We included in our study 51 SNPs in *ATM*, 14 SNPs in *CHEK2* and 13 SNPs in *ERBB2* that were successfully genotyped in the 92 controls. All included SNPs were at least 3% in minor allele frequency and were in HWE (Table 4). Mean spacing between included SNPs was 2.9 kb, 4.0 kb and 2.8 kb in *ATM, CHEK2* and *ERBB2*, respectively (Table 4). We produced LD plots from the included SNPs in the three genes and detected strong LD across all the genes (Figures 2-4). Using the SNP dropping method [197], we found that the tagSNPs selected from the included SNPs could capture non-genotyped SNPs efficiently (Table 4).

Table 4. Summary statistics on genotyping results and SNP coverage in *ATM*, *CHEK2* and *ERBB2*.

Summary statistics	*ATM*	*CHEK2*	*ERBB2*
Number of successfully genotyped SNPs[a]	152[b]	34[c]	38[d]
Number of polymorphic SNPs	68	23	16
Number of common SNPs[e]	52	19	13
Number of SNPs deviating from HWE[f]	1	5	0
Number of SNPs included in study	51	14	13
Sequence coverage (kb)	146.2	52.0	33.9
Mean spacing between SNPs (kb)	2.9	4.0	2.8
Median spacing between SNPs (kb)	2.0	3.2	2.7
Number of common haplotypes[e,g]	6	6	8
Percentage of chromosomes accounted for by common haplotypes[g]	89	81	96
Number of tagSNPs selected	7	6	7
Average tagSNP prediction of common SNPs included in study (R^2)[e]	0.96	0.95	0.99
Average tagSNP prediction of common haplotypes (R^2)[e]	0.95	0.94	0.94
Coverage evaluation[h]			
Average prediction of dropped SNPs (R^2)	0.92	0.93	0.72
Percentage of R^2 values ≥ 0.7	92	93	70

[a] In 92 controls.
[b] Supplementary Table 2 in Paper III and Supplementary Table 1 in Paper IV.
[c] Table 2 in Paper II.
[d] Supplementary Table 3 in Paper III and Supplementary Table 2 in Paper IV.
[e] Common was defined as minor allele frequency ≥ 0.03 (SNPs) or haplotype frequency ≥ 0.03.
[f] $P < 0.01$.
[g] Haplotypes were reconstructed from the SNPs included in the study.
[h] SNP dropping method by Weale et al. [197].

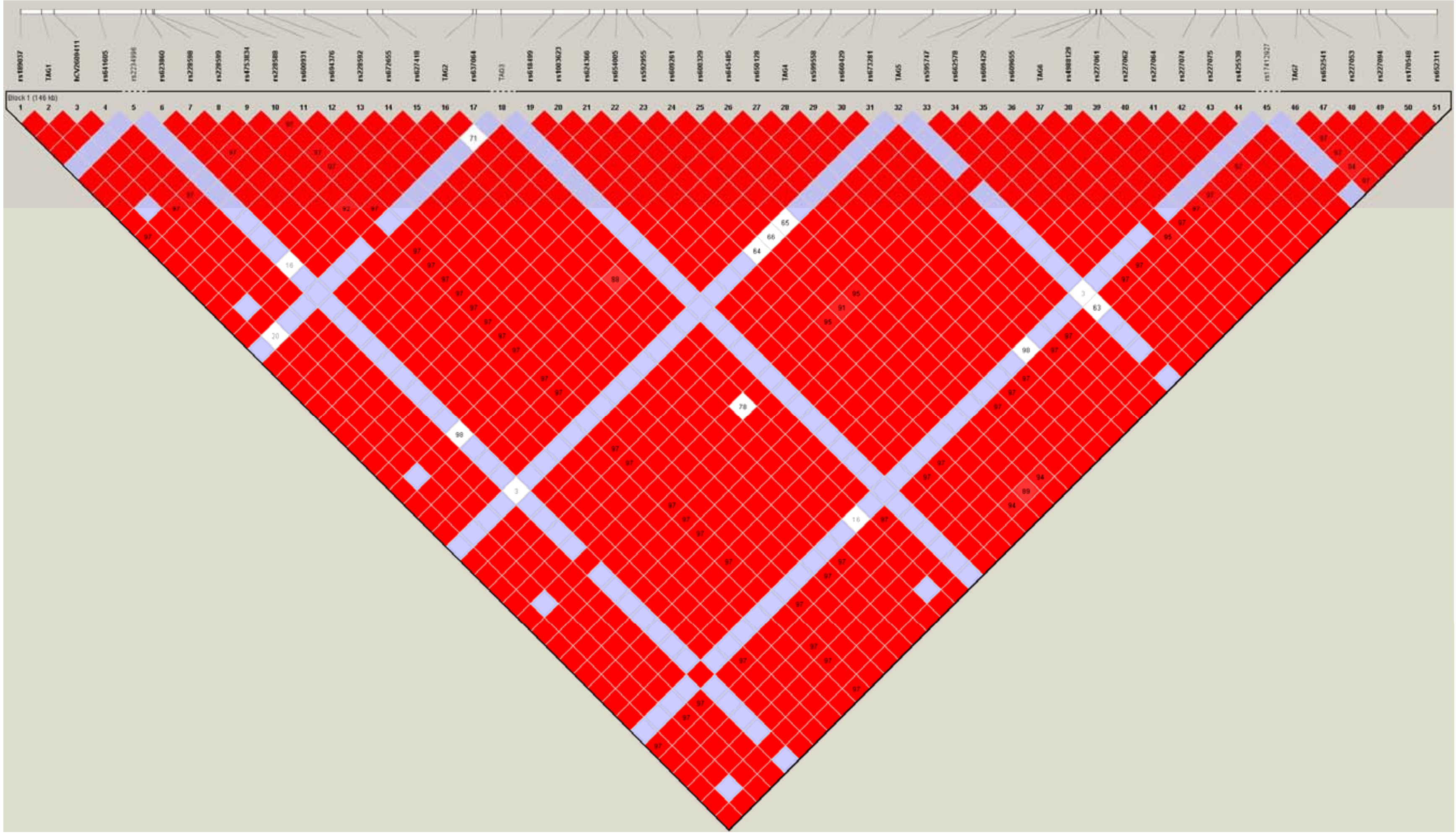

Figure 2. LD plot of 51 SNPs in *ATM* genotyped in 92 controls and included in the study. Red: D′=1 and LOD ≥ 2. Blue: D′=1 and LOD < 2. White: D′< 1 and LOD < 2.

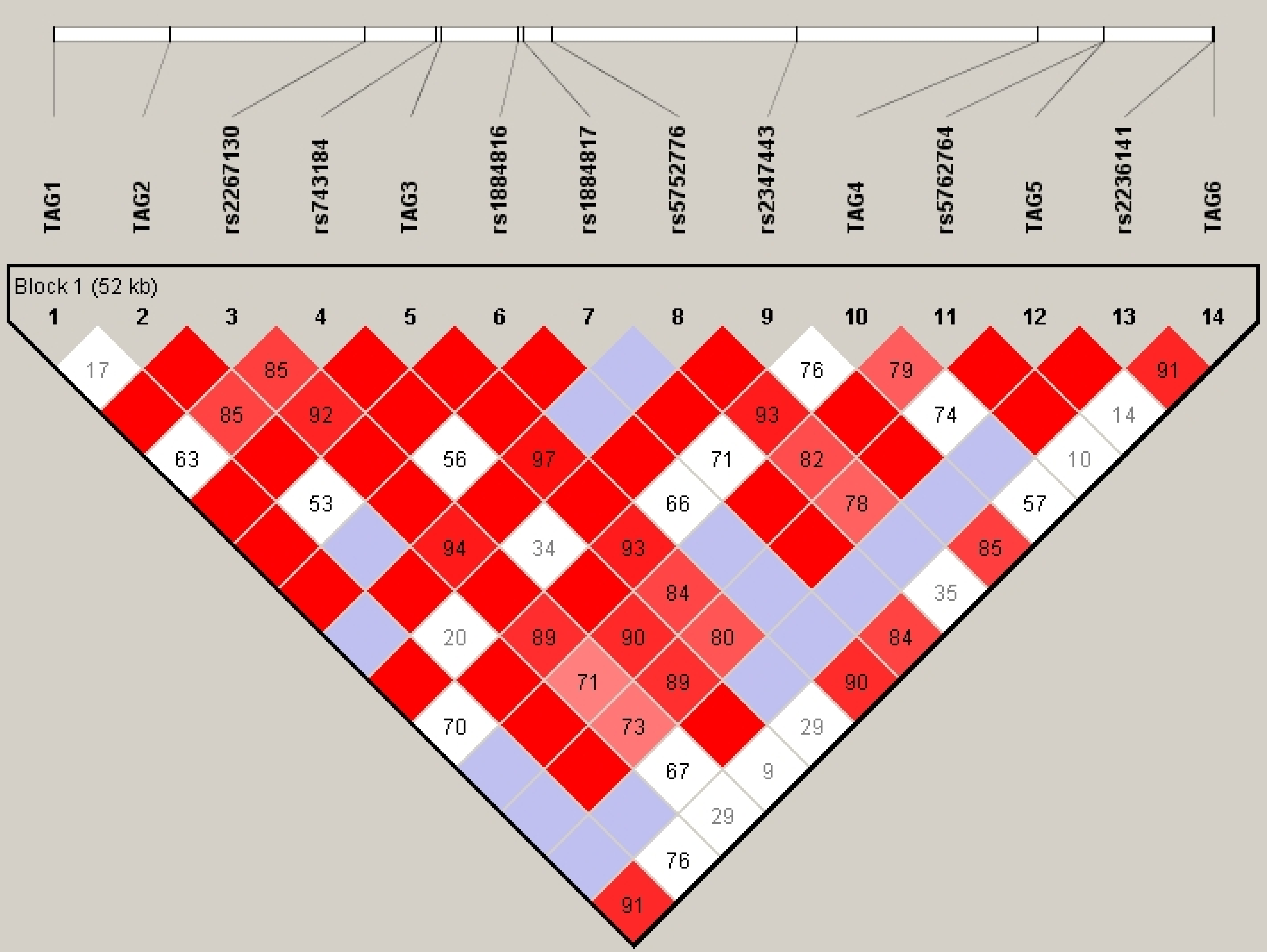

Figure 3. LD plot of 14 SNPs in *CHEK2* genotyped in 92 controls and included in the study. Red: D′=1 and LOD ≥ 2. Blue: D′=1 and LOD < 2. White: D′< 1 and LOD < 2.

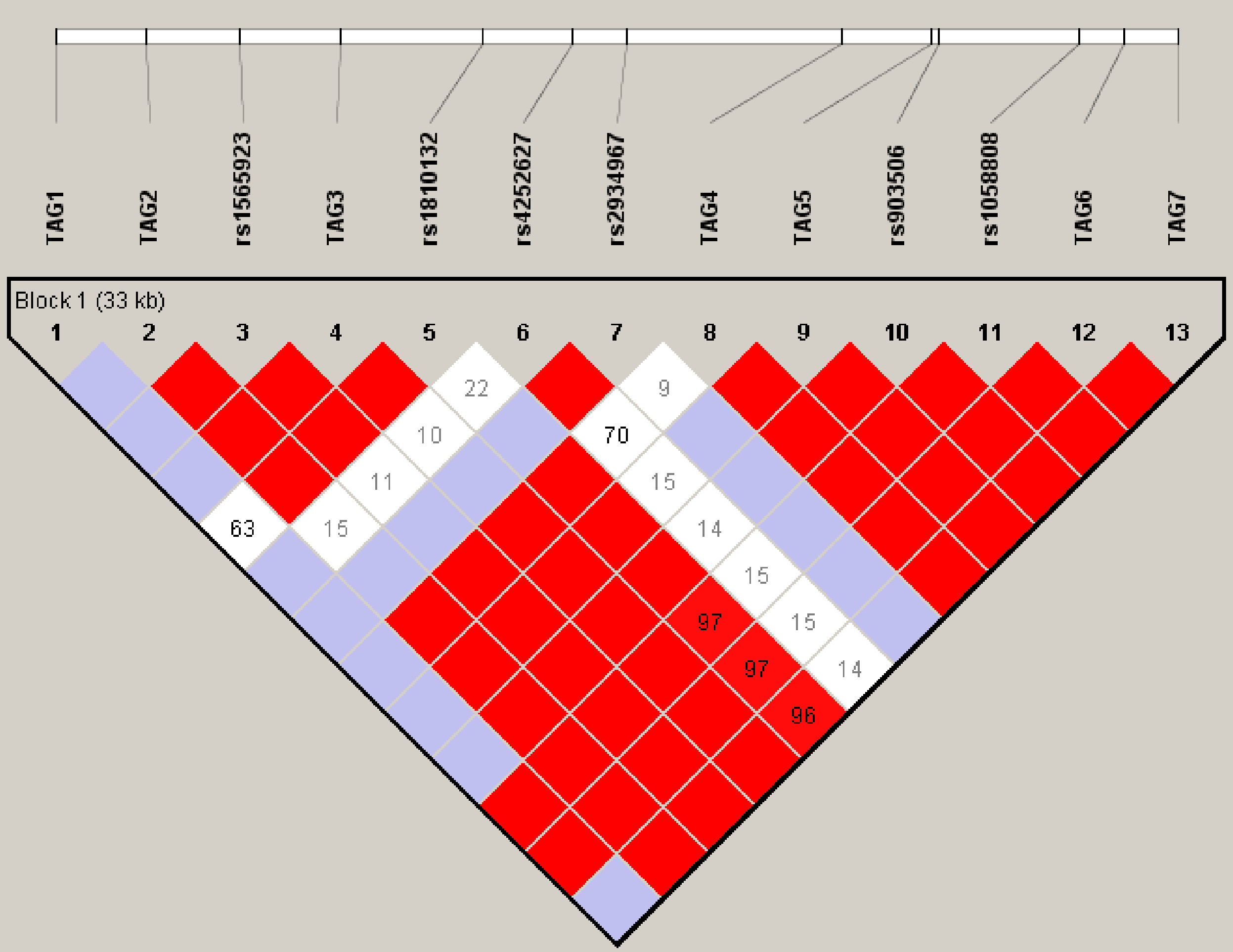

Figure 4. LD plot of 13 SNPs in *ERBB2* genotyped in 92 controls and included in the study. Red: D′=1 and LOD ≥ 2. Blue: D′=1 and LOD < 2. White: D′< 1 and LOD < 2.

From the included SNPs in *ATM*, *CHEK2* and *ERBB2*, we selected 7 tagSNPs in *ATM*, 6 tagSNPs in *CHEK2* and 7 tagSNPs in *ERBB2* that could predict the included SNPs and their haplotypes with an R^2 of at least 0.8. The tagSNPs were genotyped in all cases and controls, but five tagSNPs in *ATM* could not be genotyped in the cases who participated via tissue sample donation. All tagSNPs were in HWE among both breast and endometrial cancer controls and none showed a meaningful association with any of the breast or endometrial cancer risk factors. Only one of the tagSNPs – TAG5 in *ERBB2*, also named I655V – conferred an amino acid change in the protein product.

Overview of Association Analyses

Our testing strategy was to first assess the effect of the tagSNPs singly and then to estimate the effect of their haplotypes. The tagSNPs were included in the regression models assuming co-dominance of effects with homozygotes of the major allele as a reference group. When assessing the effect of the tagSNP haplotypes we included the dosages for the common haplotypes and the combined group of rare haplotypes in the regression models, with the most common haplotype as the reference. Only if the global *P*-values for the haplotype regression models in the subgroup analyses were significant did we scrutinize the haplotype-specific effects.

Breast Cancer Risk

tagSNPs

We found no effect of *ATM, CHEK2* or *ERBB2* tagSNPs on breast cancer risk (Table 5), which was not altered after restricting the analyses to the randomly selected cases and controls or conditioning the *ATM* and *ERBB2* analyses on menopausal hormone use or diabetes mellitus.

Table 5. Characteristics of the tagSNPs in *ATM, CHEK2* and *ERBB2* and their association with breast cancer risk.

SNP ID	dbSNP name	Alleles[a]	Number of cases/controls	Minor allele frequency[b]	HWE *P*-value[b]	OR (95% CI)[c]
ATM						
TAG1[d]	rs4987886	A/T	1220 / 1440	0.06	0.83	1.05 (0.84-1.31)
TAG2[d]	rs3092991	A/G	1119 / 1318	0.14	0.04	1.08 (0.93-1.27)
TAG3[d]	rs1800057	C/G	1144 / 1346	0.03	0.30	0.89 (0.63-1.26)
TAG4	rs1801516	G/A	1538 / 1500	0.15	0.34	1.08 (0.94-1.24)
TAG5	rs17107917	C/G	1546 / 1493	0.04	0.05	1.00 (0.79-1.28)
TAG6[d]	rs227060	C/T	1152 / 1350	0.28	0.66	0.99 (0.88-1.12)
TAG7[d]	rs664143	C/T	1227 / 1408	0.48	0.20	1.01 (0.91-1.13)
CHEK2						
TAG1	rs8135424	G/A	1539 / 1478	0.13	0.34	1.15 (0.99-1.34)
TAG2	rs5762749	C/G	1516 / 1472	0.35	0.80	1.00 (0.89-1.12)
TAG3	rs743185	C/T	1547 / 1491	0.12	0.63	1.05 (0.90-1.24)
TAG4	rs738722	C/T	1501 / 1444	0.25	0.38	1.00 (0.89-1.13)
TAG5	rs5762765	G/C	1510 / 1456	0.38	0.10	0.99 (0.88-1.10)
TAG6	rs2236142	C/G	1541 / 1471	0.31	0.85	1.09 (0.97-1.22)
ERBB2						
TAG1	rs2643195	G/A	1494 / 1458	0.32	0.16	0.98 (0.88-1.09)
TAG2	rs4252596	G/A	1530 / 1481	0.13	0.08	0.90 (0.78-1.05)
TAG3	rs2952155	C/T	1459 / 1407	0.25	0.84	0.99 (0.88-1.11)
TAG4	rs2952156	G/A	1546 / 1481	0.32	0.76	0.97 (0.87-1.09)
TAG5[e]	rs1801200	A/G	1548 / 1485	0.26	0.39	1.01 (0.91-1.13)
TAG6	rs4252665	C/T	1527 / 1486	0.05	0.94	1.00 (0.80-1.25)
TAG7	rs3809717	G/T	1532 / 1478	0.31	0.60	1.01 (0.90-1.12)

[a] Major alleles given first, minor alleles second.
[b] In all breast cancer controls.
[c] Odds ratios are assessed assuming co-dominance and show the increase/decrease in breast cancer risk with each addition of the rare allele. *ATM* and *ERBB2* analyses were adjusted for age. *CHEK2* analyses were also adjusted for menopausal hormone use and diabetes mellitus.
[d] Not genotyped in the cases who participated via tissue sample donation.
[e] Also named I655V.

Haplotypes

Compared to the most common haplotype, the rare haplotypes in *CHEK2* appeared to increase breast cancer risk (Table 6). The association did not however carry over to the global test. After excluding the *1100delC* carriers (n=28) from the analysis, this association decreased (OR 1.20; 95% CI, 0.98–1.47), whilst the odds ratios for the common haplotypes remained unchanged. Consideration only of breast cancer cases eventually diagnosed with a second breast cancer (n=72) did not provide any convincing association. These findings remained unaltered after restricting the analyses to the randomly selected cases and controls or when we compared carriers to noncarriers of each haplotype, instead of using haplotype 1 as reference.

Table 6. Common tagSNP haplotypes in *ATM, CHEK2* and *ERBB2* in relation to breast cancer risk.

		Haplotype proportions		
Haplotype no.	Haplotypes	Cases	Controls	OR (95% CI)[b]
ATM		(n = 1574[a])	(n = 1513[a])	
1	AACGCCT	0.414	0.408	1.00 (Reference)
2	AACGCTC	0.231	0.231	0.99 (0.86-1.13)
3	AGCACCC	0.150	0.139	1.06 (0.91-1.24)
4	AACGCCC	0.062	0.076	0.81 (0.64-1.02)
5	TACGCCT	0.064	0.061	1.03 (0.81-1.30)
6	AACGGTC	0.043	0.043	0.97 (0.75-1.25)
	Rare[c]	0.037	0.042	0.88 (0.66-1.16)
Global *P*-value[d]				0.50
CHEK2		(n = 1571[a])	(n = 1513[a])	
1	GCCCCC	0.223	0.241	1.00 (Reference)
2	GGCTGC	0.231	0.230	1.07 (0.92-1.26)
3	GCCCCG	0.140	0.129	1.13 (0.93-1.37)
4	ACCCGC	0.113	0.104	1.20 (0.98-1.46)
5	GCTCGG	0.089	0.088	1.10 (0.88-1.36)
6	GGCCGC	0.052	0.060	0.94 (0.72-1.24)
7	GCCCGC	0.027	0.034	0.87 (0.61-1.26)
	Rare[e]	0.125	0.114	1.24 (1.02-1.51)
Global *P*-value[d]				0.19
ERBB2		(n = 1579[a])	(n = 1516[a])	
1	GGCGACT	0.296	0.295	1.00 (Reference)
2	AGTAACG	0.166	0.165	1.01 (0.86-1.18)
3	GGCGGCG	0.135	0.128	1.04 (0.88-1.23)
4	GACGACG	0.116	0.128	0.91 (0.77-1.08)
5	AGTAGCG	0.075	0.077	0.97 (0.79-1.20)
6	AGCAACG	0.068	0.071	0.94 (0.76-1.17)
7	GGCGACG	0.079	0.069	1.14 (0.91-1.41)
8	GGCGGTG	0.048	0.051	0.94 (0.74-1.20)
	Rare[f]	0.018	0.015	1.18 (0.77-1.81)
Global *P*-value[d]				0.76

[a] Information on at least one tagSNP

[b] *ATM* and *ERBB2* analyses were adjusted for age. *CHEK2* analyses were also adjusted for menopausal hormone use and diabetes mellitus.

[c] 11 rare haplotypes combined. Each haplotype has frequency below 3% among the controls.

[d] Likelihood ratio test.

[e] 19 rare haplotypes combined. Each haplotype has frequency below 3% among the controls.

[f] 19 rare haplotypes combined. Each haplotype has frequency below 3% among the controls.

We stratified the haplotype analyses by several breast cancer risk factors: Age at menarche and menopause, first-degree family history, body mass index, age at first birth, menopausal hormone use, and parity. None of the stratified analyses yielded a significant (α=0.05) global test of association.

Rare Variants

The *1100delC* mutation, genotyped in 1,510 cases and 1,334 controls, was rare in our Swedish population, with a frequency of 0.4% in the controls (HWE, *P* = 0.89).

The deletion was slightly more common in the cases (0.7%) than in the controls, and the corresponding age- and sampling scheme-adjusted odds ratio for carriers versus noncarriers was 2.26 (95% CI, 0.99–5.15) (Table 7). The *1100delC* was exclusively carried on rare haplotypes, which may explain the marginally significant association between the group of rare haplotypes in *CHEK2* and breast cancer risk (Table 6).

We genotyped two missense mutations in *ATM* in the complete sample set: 4258 C→T (rs1800058, L1420F) and 2572 T→C (rs1800056, F858L). Neither mutation deviated significantly from HWE in controls. They were both rare in our population, with minor allele frequency of 1.9% for 4258 C→T and 1.4% for 2527 T→C in the controls. When exploring the change in breast cancer risk with each addition of the rare allele compared to non-carriers (assuming co-dominance), we found elevated risk – though not significantly – for the 4258 C→T (OR 1.36, 95% CI 0.91-2.04), but no association emerged between the 2527 T→C and breast cancer risk (OR 1.05, 95% CI 0.65-1.71).

Breast Cancer Survival

We estimated the risk of breast cancer death associated with the tagSNPs (Table 8) in *ATM*, *CHEK2* and *ERBB2*, and their haplotypes (Table 9). We found decreased risk of breast cancer death associated with each addition of the rare allele of TAG2 in *CHEK2* (P = 0.026) as well as elevated risk of *CHEK2* TAG6 (P = 0.03), compared to homozygotes of the common allele for each variant (Table 8). The associations did not however withstand multiple

Table 7. Association of the *CHEK2*1100delC* with breast cancer risk.

1100delC genotype	Number of cases/controls	OR (95% CI)[a]
C/C	1490/1326	1 (Reference)
C/-	19/8	2.13 (0.92-4.89)
-/-	1/0	---
C/- and -/-	20/8	2.26 (0.99-5.15)

[a] Analyses were conditioned on age, menopausal hormone use and diabetes mellitus.

Table 8. Association of the tagSNPs in *ATM*, *CHEK2* and *ERBB2* with breast cancer survival.

SNP ID	Br.ca. deaths /person-years[a]	HR (95% CI)[a,b]
ATM		
TAG1[c]	70 / 10,660	0.86 (0.42-1.76)
TAG2[c]	65 / 9,717	1.12 (0.71-1.77)
TAG3[c]	66 / 9,958	0.62 (0.16-2.46)
TAG4	185 / 12,421	0.99 (0.75-1.30)
TAG5	192 / 12,466	1.01 (0.63-1.64)
TAG6[c]	68 / 10,004	0.77 (0.52-1.14)
TAG7[c]	73 / 10,684	1.24 (0.89-1.73)
CHEK2		
TAG1	197 / 12,367	1.08 (0.83-1.41)
TAG2	191 / 12,172	0.78 (0.62-0.97)
TAG3	192 / 12,471	1.21 (0.91-1.62)
TAG4	187 / 12,083	0.72 (0.56-0.93)
TAG5	192 / 12,118	1.15 (0.94-1.41)
TAG6	192 / 12,398	1.25 (1.02-1.54)
ERBB2		
TAG1	186 / 12,052	1.09 (0.88-1.35)
TAG2	189 / 12,330	0.95 (0.70-1.30)
TAG3	184 / 11,666	1.02 (0.80-1.29)
TAG4	194 / 12,445	1.11 (0.89-1.37)
TAG5[e]	193 / 12,463	1.00 (0.80-1.25)
TAG6	182 / 12,401	0.76 (0.45-1.27)
TAG7	194 / 12,310	0.96 (0.77-1.20)

[a] Among breast cancer cases.
[b] Hazard ratios are assessed assuming co-dominance and show the increase/decrease in risk of breast cancer death with each addition of the minor allele compared to homozygotes of the major allele.
[c] Not genotyped in the cases who participated via tissue sample donation.
[e] Also named I655V.

testing correction. Carriers of haplotype 2 in *CHEK2* appeared to have decreased risk of breast cancer death (*P* = 0.038), compared to haplotype 1 carriers, whilst *ERBB2* rare haplotype carriers seemed to have increased risk (*P* = 0.009). Neither association carried over to the global test (*P* = 0.15 and *P* = 0.45, respectively) (Table 9).

Table 9. tagSNP haplotypes in *ATM, CHEK2* and *ERBB2* in relation to breast cancer survival.

Haplotype number	Haplotypes	Haplotype proportions (cases)	HR (95% CI)
ATM		n = 1574[a]	
1	AACGCCT	0.414	1.00 (Reference)
2	AACGCTC	0.231	0.85 (0.65-1.13)
3	AGCACCC	0.150	0.89 (0.66-1.21)
4	AACGCCC	0.062	0.86 (0.52-1.44)
5	TACGCCT	0.064	0.88 (0.55-1.40)
6	AACGGTC	0.043	0.95 (0.57-1.57)
	Rare[b]	0.037	0.95 (0.53-1.68)
Global *P*-value[c]			0.95
CHEK2		n = 1571[a]	
1	GCCCCC	0.223	1.00 (Reference)
2	GGCTGC	0.231	0.72 (0.52-0.98)
3	GCCCCG	0.140	1.02 (0.72-1.43)
4	ACCCGC	0.113	1.00 (0.70-1.41)
5	GCTCGG	0.089	1.08 (0.73-1.57)
6	GGCCGC	0.052	0.61 (0.34-1.10)
7	GCCCGC	0.027	0.60 (0.26-1.41)
	Rare[d]	0.125	0.95 (0.67-1.35)
Global *P*-value[c]			0.15
ERBB2		n = 1579[a]	
1	GGCGACT	0.296	1.00 (Reference)
2	AGTAACG	0.166	0.98 (0.71-1.34)
3	GGCGGCG	0.135	1.01 (0.73-1.40)
4	GACGACG	0.116	0.98 (0.69-1.39)
5	AGTAGCG	0.075	1.21 (0.81-1.81)
6	AGCAACG	0.068	1.16 (0.76-1.77)
7	GGCGACG	0.079	1.06 (0.69-1.63)
8	GGCGGTG	0.048	0.81 (0.47-1.39)
	Rare[e]	0.018	2.21 (1.22-4.02)
Global *P*-value[c]			0.45

[a] Information on at least one tagSNP
[b] 11 rare haplotypes combined. Each haplotype has frequency below 3% among the controls.
[c] Likelihood ratio test.
[d] 19 rare haplotypes combined. Each haplotype has frequency below 3% among the controls.
[e] 19 rare haplotypes combined. Each haplotype has frequency below 3% among the controls.

Breast Tumour Characteristics

Breast cancer cases were divided in groups according to their tumour characteristics (see 'Subjects', 'Breast Tumour Characteristics and Follow-up' above) and each group was contrasted against all controls. None of the global *P*-values of the haplotype logistic regression models reached significance, which indicates that after taking into account the number of tests performed, none of the individual haplotypes affected the risk of developing tumours with certain characteristics.

We could not perform a meaningful analysis regarding the association between the rare variants in *CHEK2* and *ATM* and breast cancer characteristics or survival due to very small sample sizes.

Endometrial Cancer Risk

tagSNPs

When assessing the change in risk with each addition of the rare allele compared to non-carriers, we found TAG1 in *CHEK2* to be associated with increased endometrial cancer risk (*P* = 0.01, Table 10), but multiple testing adjustment rendered the association non-significant (*P* = 0.23). Restricting the analysis to include only endometroid tumours yielded a stronger association with an odds ratio of 1.28 (95% CI 1.07-1.54, *P* = 0.007).

Table 10. Characteristics of the tagSNPs genotyped in *ATM, CHEK2* and *ERBB2* and their association with endometrial cancer risk.

SNP ID	Minor allele frequency[a]	HWE *P*-value[a]	Number of cases/controls	OR (95% CI)[b]
ATM				
TAG1[c]	0.06	0.69	552/1467	0.97 (0.73-1.30)
TAG2[c]	0.14	0.01	501/1361	1.14 (0.93-1.38)
TAG3[c]	0.03	0.39	523/1390	0.84 (0.54-1.30)
TAG4	0.14	0.18	694/1547	1.09 (0.91-1.30)
TAG5	0.04	0.08	690/1539	1.15 (0.87-1.54)
TAG6[c]	0.28	0.90	521/1393	1.02 (0.87-1.19)
TAG7[c]	0.48	0.55	546/1450	1.02 (0.89-1.18)
CHEK2				
TAG1	0.13	0.94	683/1524	1.26 (1.06-1.51)
TAG2	0.35	0.95	672/1516	0.93 (0.82-1.07)
TAG3	0.12	0.51	682/1541	0.93 (0.76-1.14)
TAG4	0.26	0.66	663/1493	1.03 (0.89-1.19)
TAG5	0.39	0.07	671/1500	0.96 (0.84-1.10)
TAG6	0.31	0.94	682/1521	1.02 (0.89-1.17)
ERBB2				
TAG1	0.32	0.21	671/1503	0.91 (0.80-1.05)
TAG2	0.13	0.05	687/1524	0.93 (0.77-1.13)
TAG3	0.26	0.90	657/1454	0.93 (0.80-1.09)
TAG4	0.32	0.81	686/1530	0.91 (0.79-1.05)
TAG5[e]	0.26	0.50	691/1531	0.95 (0.83-1.10)
TAG6	0.05	0.58	690/1534	0.95 (0.71-1.28)
TAG7	0.31	0.81	682/1524	1.03 (0.90-1.18)

[a] In all endometrial cancer controls.
[b] Odds ratios are assessed assuming co-dominance and show the increase/decrease in endometrial cancer risk with each addition of the rare allele. Analyses were adjusted for age (5 year age-groups).
[c] Not genotyped in cases who participated via tissue sample donation.
[e] Also named I655V.

When we explored individual genotype risks for *CHEK2* TAG1, the increased risk appeared to be confined to homozygous carriers of the rare allele (AA) (Table 11). Compared with GG carriers, the risk in AA carriers was 2.11 (*P* = 0.012) for all tumours and 2.29 (*P* = 0.005) for the endometroid tumours (Table 11). Conditioning on the selection variables (menopausal hormone use and diabetes mellitus) or restricting the analyses to the randomly selected controls did not alter the results.

Table 11. Association of TAG1 in *CHEK2* with endometrial cancer risk overall and restricted to endometroid tumours.

	All cancers		Endometroid tumours	
TAG1 *CHEK2*	No. of cases/controls	OR (95% CI)[a]	No. of cases/controls	OR (95% CI)[a]
GG	490/1156	1.00 (reference)	453/1156	1.00 (reference)
GA	170/343	1.18 (0.95-1.46)	157/343	1.18 (0.95-1.47)
AA	23/25	2.11 (1.18-3.77)	23/25	2.29 (1.28-4.08)

[a] Adjusted for age in 5-year age-groups.

Haplotypes

When we assessed the association of the tagSNP haplotypes in *ATM, CHEK2* and *ERBB2* in relation to endometrial cancer risk, haplotype 4 in *ATM* and haplotype 4 in *CHEK2* appeared to affect the risk (*P* = 0.028 and *P* = 0.017, respectively) compared to haplotype 1 in each gene (data now shown). The associations did not, however, carry over to the global tests for either gene (*P* = 0.20 and *P* = 0.24, respectively). These results were unaffected after conditioning on the selection variables (menopausal hormone use and diabetes mellitus), or restricting the analyses to the randomly selected controls.

Global likelihood ratio test *P*-values for association between *ATM*, *CHEK2* and *ERBB2* haplotypes and endometrial cancer risk, restricted to certain tumour subtypes or stratified by cancer risk factors are shown in Table 12. We did not perform tests within the subgroups of medium potency estrogen only or in combination with progestin since the low numbers might have affected the reliability of the global *P*-values. *ATM* haplotypes appeared to affect endometrial cancer risk among women who delivered their last child over 33 years of age (global *P* = 0.027, Table 12). Haplotype 5 showed a borderline non-significant association (*P* = 0.053) compared to haplotype 1 in this group of women, but the likelihood ratio test for interaction between age at last birth and haplotype 5 in *ATM* was not statistically significant (*P* = 0.08). A stronger association emerged between endometrial cancer risk and *ATM* haplotypes in non-smokers (global *P* = 0.009), but became non-significant after multiple testing adjustment (*P* = 0.32). This association was driven by haplotype 4 in *ATM* (*P* = 0.002), which decreased the risk of endometrial cancer (OR 0.50, 95% CI 0.32-0.77) compared to haplotype 1 (Table 13). When we compared carriers of haplotype 4 with non-carriers, the association was slightly stronger (OR 0.48, 95% CI 0.31-0.73, *P* = 0.0007), and the test of interaction indicated that the effect of haplotype 4 in *ATM* on endometrial cancer risk depended on smoking status (*P* = 0.0037).

CHEK2 haplotypes were associated with endometrial cancer risk among women with menopause below 49 years of age (global *P* = 0.034, Table 12). In this group of women, all haplotypes in *CHEK2* appeared to increase endometrial cancer risk when compared with haplotype 1 (Table 13). However, when we compared each haplotype with non-carriers of the respective haplotype, only haplotype 1 affected endometrial cancer risk among these women (OR 0.50, 95% CI 0.33-0.75, *P* = 0.0009). The risk related to haplotype 1 increased with increasing age at menopause (49-52 years: OR = 0.88, *P* = 0.30; >52 years: OR = 1.17, *P* = 0.31) and the test for interaction between age at menopause and haplotype 1 in *CHEK2* was statistically significant (*P* = 0.007).

Table 12. Global *P*-values for the association of *ATM, CHEK2* and *ERBB2* tagSNP haplotypes with endometrial cancer risk, restricted to tumour subtypes and stratified by endometrial cancer risk factors.

	ATM	*CHEK2*	*ERBB2*
Characteristic	Global *P*-value[a]	Global *P*-value[b]	Global *P*-value[c]
Endometroid cancers	0.261	0.203	0.493
Grade I	0.300	0.912	0.536
Grade II	0.423	0.156	0.865
Grade III	0.842	0.286	0.413
Myometrial invasion[d]			
No	0.418	0.226	0.778
Yes	0.671	0.542	0.057
Age at menopause (years)			
<49	0.435	0.034	0.508
49-52	0.427	0.050	0.866
>52	0.080	0.807	0.248
Age at last birth (years)			
≤26	0.723	0.391	0.311
27-33	0.153	0.693	0.810
≥34	0.027	0.284	0.486
Parity			
Nulliparous	0.561	0.682	0.164
1 child	0.659	0.253	0.789
2 children	0.801	0.638	0.559
≥3 children	0.195	0.457	0.641
Body mass index (kg/m2)			
<25	0.266	0.690	0.508
25-<28	0.306	0.682	0.128
≥28	0.372	0.078	0.212
Regular smoking for at least 1 year			
No	0.009	0.209	0.279
Yes	0.261	0.847	0.344
Family history[e]			
No	0.066	0.329	0.365
Yes	0.431	0.369	0.109
Combined oral contraceptives[f]			
Never	0.135	0.221	0.795
Ever	0.684	0.585	0.352
Low potency estrogen use[g]			
Never	0.240	0.320	0.135
Ever	0.815	0.392	0.756
Self-reported diabetes mellitus			
No	0.401	0.455	0.585
Yes	0.304	0.738	0.155

[a] Likelihood ratio test with 6 degrees of freedom. Models include 5 common haplotypes and the 10 rare haplotypes combined into a single variable. The most common haplotype is the reference.

[b] Likelihood ratio test with 6 degrees of freedom. Models include 5 common haplotypes and the 19 rare haplotypes combined into a single variable. The most common haplotype is the reference.

[c] Likelihood ratio test with 8 degrees of freedom. Models include 7 common haplotypes and the 16 rare haplotypes combined into a single variable. The most common haplotype is the reference.

[d] No: No invasion or <50% of the myometrum. Yes: Invasion through ≥ 50% of the myometrium or through the serosa.

[e] At least one 1st degree relative with endometrial cancer

[f] Estrogens and progestins given concurrently in a monthly cycle.

[g] Oestriol or oestradiol of low dose. Not exclusive use.

Table 13. The association of *ATM, CHEK2* and *ERBB2* tagSNP haplotypes with endometrial cancer risk stratified by smoking history and age at menopause.

	ATM	*CHEK2*	*ERBB2*
Smoking (no)[d]	OR (95% CI)[a]	OR (95% CI)[b]	OR (95% CI)[c]
haplotype 1	1 (Reference)	1 (Reference)	1 (Reference)
haplotype 2	0.97 (0.78-1.22)	1.14 (0.89-1.47)	0.94 (0.73-1.22)
haplotype 3	1.14 (0.89-1.46)	1.16 (0.85-1.58)	0.93 (0.70-1.23)
haplotype 4	0.50 (0.32-0.77)	1.49 (1.10-2.01)	1.22 (0.93-1.60)
haplotype 5	1.17 (0.82-1.67)	1.00 (0.72-1.39)	0.78 (0.54-1.12)
haplotype 6	1.45 (0.99-2.12)	1.02 (0.67-1.55)	0.95 (0.68-1.33)
haplotype 7	---	---	1.27 (0.91-1.77)
haplotype 8	---	---	1.20 (0.80-1.78)
Rare	0.98 (0.63-1.53)	1.16 (0.87-1.55)	0.78 (0.37-1.64)
Global *P*-value[e]	0.01	0.21	0.28
Smoking (yes)[d]			
haplotype 1	1 (Reference)	1 (Reference)	1 (Reference)
haplotype 2	0.92 (0.70-1.22)	0.93 (0.68-1.28)	1.12 (0.82-1.53)
haplotype 3	0.92 (0.67-1.25)	1.14 (0.78-1.66)	0.85 (0.60-1.22)
haplotype 4	1.05 (0.68-1.63)	1.12 (0.78-1.62)	0.84 (0.59-1.20)
haplotype 5	0.68 (0.42-1.12)	1.02 (0.64-1.65)	0.78 (0.50-1.20)
haplotype 6	0.77 (0.46-1.30)	0.78 (0.43-1.43)	0.74 (0.46-1.21)
haplotype 7	---	---	0.95 (0.62-1.46)
haplotype 8	---	---	0.58 (0.33-1.00)
Rare	0.44 (0.21-0.9)	1.05 (0.75-1.47)	1.47 (0.56-3.88)
Global *P*-value[e]	0.26	0.85	0.34
Age at menopause <49 years			
haplotype 1	1 (Reference)	1 (Reference)	1 (Reference)
haplotype 2	0.82 (0.54-1.22)	2.02 (1.25-3.28)	0.65 (0.41-1.04)
haplotype 3	0.73 (0.47-1.15)	2.53 (1.42-4.52)	0.56 (0.33-0.97)
haplotype 4	0.71 (0.33-1.53)	2.17 (1.23-3.81)	0.91 (0.57-1.44)
haplotype 5	1.22 (0.62-2.41)	1.46 (0.75-2.85)	0.67 (0.36-1.27)
haplotype 6	1.20 (0.61-2.38)	1.84 (0.85-4.00)	0.69 (0.36-1.30)
haplotype 7	---	---	0.78 (0.41-1.51)
haplotype 8	---	---	0.64 (0.31-1.30)
Rare	1.51 (0.72-3.17)	1.96 (1.13-3.39)	1.16 (0.38-3.55)
Global *P*-value[e]	0.43	0.03	0.51
Age at menopause 49-52 years			
haplotype 1	1 (Reference)	1 (Reference)	1 (Reference)
haplotype 2	1.04 (0.80-1.37)	1.03 (0.75-1.40)	1.06 (0.78-1.43)
haplotype 3	1.26 (0.93-1.71)	1.04 (0.71-1.52)	0.97 (0.69-1.36)
haplotype 4	0.95 (0.61-1.48)	1.77 (1.24-2.53)	1.10 (0.78-1.53)
haplotype 5	0.83 (0.52-1.32)	1.17 (0.77-1.76)	0.76 (0.49-1.17)
haplotype 6	1.02 (0.61-1.68)	1.06 (0.63-1.78)	0.86 (0.55-1.33)
haplotype 7	---	---	0.84 (0.56-1.28)
haplotype 8	---	---	0.82 (0.49-1.35)
Rare	0.66 (0.37-1.19)	1.03 (0.73-1.46)	1.28 (0.50-3.27)
Global *P*-value[e]	0.43	0.05	0.87
Age at menopause >52 years			
haplotype 1	1 (Reference)	1 (Reference)	1 (Reference)
haplotype 2	1.24 (0.89-1.74)	0.82 (0.56-1.19)	1.34 (0.93-1.95)
haplotype 3	1.19 (0.82-1.72)	0.87 (0.55-1.38)	0.90 (0.58-1.38)
haplotype 4	0.45 (0.23-0.87)	0.78 (0.49-1.24)	1.05 (0.69-1.59)
haplotype 5	1.23 (0.71-2.13)	0.77 (0.46-1.29)	0.88 (0.52-1.51)
haplotype 6	1.33 (0.78-2.29)	0.66 (0.34-1.29)	0.92 (0.55-1.56)
haplotype 7	---	---	1.71 (1.04-2.82)
haplotype 8	---	---	1.35 (0.73-2.51)

	ATM	*CHEK2*	*ERBB2*
Rare	0.68 (0.31-1.50)	0.98 (0.66-1.46)	0.57 (0.15-2.15)
Global *P*-value[e]	0.08	0.81	0.25

[a] Models include 5 common haplotypes and the 10 rare haplotypes combined into a single variable. The most common haplotype is the reference.
[b] Models include 5 common haplotypes and the 19 rare haplotypes combined into a single variable. The most common haplotype is the reference.
[c] Models include 7 common haplotypes and the 16 rare haplotypes combined into a single variable. The most common haplotype is the reference.
[d] Regular smoking for at least one year or ever having smoked over 100 cigarettes.
[e] Likelyhood ratio test.

Rare Variants

We genotyped the non-synonymous variants rs1800056 (2572 T→C, F858L) and rs1800058 (4258 C→T, L1420F) in the *ATM* gene and the *1100delC* deletion in the *CHEK2* gene in our endometrial cancer sample set. All variants were very rare in our population with minor allele frequencies of 1.6%, 1.7% and 0.4% among the controls respectively. We found no association between any of the three variants and endometrial cancer risk.

DISCUSSION

Study Design

Our study was a population-based case-control study. Case-control studies can be thought of as reverse cohort studies. In a cohort study, a population is defined, exposure information determined, and the population is followed to see how many develop the disease and whether they are more or less frequently exposed than those who do not develop the disease. In a population-based case-control study, a population is defined, all cases are identified, controls are selected, and the exposure distribution – which often happened many years earlier – is determined in both cases and controls. The controls should be sampled from the source population that gave rise to the cases and should be sampled independently of the exposure such that the exposure distribution in the controls represents that of the source population [207]. Hence, if the selection of the controls does not depend on the exposure distribution in any way and measurement of the exposure will not be different between cases and controls, the case-control study will be a valid approximation of a cohort study.

Case-control studies are often preferred to cohort studies when the disease is rare, has a long induction and latent period and/or if expensive laboratory tests are required to be carried out on biological samples obtained from the cohort members. For example, if researchers design a case-control study instead of a cohort study they do not have to wait for decades for only few cases to develop the disease and do not have to perform expensive laboratory tests on thousands of biological samples. We studied breast and endometrial cancer, which are both rare diseases with a long induction period. Furthermore, we collected biological samples from all participants and performed numerous expensive laboratory tests. Hence, the case-control design we applied was obviously to our advantage.

It is believed that case-control studies are more prone to bias such as selection bias and recall bias than cohort studies since the study base (the population) is often not well defined and the exposure information is collected retrospectively [207]. Furthermore, it can be difficult to establish a clear temporal relationship between the exposure and disease because of the retrospective nature of the data. Case-control studies were thus thought of in the past as being less valid than cohort studies. This view has changed in the last couple of decades since researchers realized that well-designed case-control studies can be just as efficient a way to learn about the relationship between an exposure and disease as cohort studies.

In order to determine the internal validity of our study and subsequently to determine causation, I discuss below some of the main considerations for case-control studies as well as genetic association studies and relate these factors to our study.

Validity

Internal validity of a study must be established before the observed results can be deemed causal and before the results can be generalized to other populations. Only if bias, confounding and random error have been carefully considered and found to be negligible can the investigator conclude that the study is valid and the association true [207].

Selection Bias

Control Selection Bias

A well defined study base is the foundation of all case-controls studies. A problem with this definition can lead to selection bias. This can happen when the investigators do not use the same criteria to select cases and controls. In hospital-based case-control studies for example, when the cases and controls – with an unrelated disease – are identified from hospital records, the investigators have to make sure that the illness of the controls has the same referral pattern to the health care facility as that of the cases [207]. This can be very difficult to determine particularly since the study base is often unknown. Population-based case-control studies like ours circumvent this problem by defining the study base prior to selecting cases and controls, thus ensuring that the controls are selected from the population that gave rise to the cases.

Self-selection Bias

Refusal to participate in the study or non-response by the eligible cases and controls can lead to selection bias if this non-participation is different between cases and controls and is related to the exposure [207]. Selection bias of this type could be a concern in our study since non-participation was related to severe disease or death. This non-participation was related to case-control status since the cases were more likely to become seriously ill and it might have been related to our exposure; genetic variation.

We sought to obtain tissue samples from the deceased cases and those cases that had declined donation of a blood sample, and were able to obtain the majority of the samples requested. The relative minor lack of tissue accessibility is unlikely to be related to our exposure – *CYP17*, *ATM*, *CHEK2* or *ERBB2* genetic variation – as it depended on the inability of the respective pathology department to retrieve the samples. Furthermore, genotype frequencies of the tagSNPs in *ATM*, *CHEK2* and *ERBB2* did not differ between blood and tissue samples, which suggests that the exposure was not related to non-participation and therefore that selection bias was negligible. Hence, we believe our main problem was lack of generalizability to women with severe breast cancer.

However, we were not able to genotype five tagSNPs in *ATM* in the tissue samples. If these five tagSNPs were in fact associated with severe disease, the association with risk of breast cancer death might have been biased towards the null in our study since we did not include all the severe cases. The fact that the results were not different when we restricted our analyses to the most severe cases among those who donated blood samples indicates that the five tagSNPs were truly not associated with severe disease.

Differential Diagnosis
If diagnosis of the disease is related to the exposure then selection bias can occur. In this situation, cases are selected on the basis of the exposure. It is almost impossible that diagnosis of breast or endometrial cancer could have been influenced by the main exposure in our study, which is germ-line genetic variation. However, users of menopausal hormone therapy tend to have more frequent mammographic screenings than non-users, which can lead to increased possibility of being diagnosed with breast cancer. Unless the genetic factors were associated with menopausal hormone therapy – which we found no evidence of – or to the tendency to seek medical care, this kind of selection bias is not a problem in our study.

Observation Bias

Recall Bias
Recall bias occurs when cases are more or less likely than controls to recall and report prior exposures [207]. The exposure in our study was genetic information detected in extracted DNA from blood or tissue, information which the participants did not have any knowledge of. It is therefore impossiple that the exposure information could have depended on the respective memory of the cases and controls. However, we had extensive questionnaire information on reproductive and lifestyle factors, which enabled us to test for gene-environment interactions. Recall of these factors could have been differential between cases and controls. Nevertheless, provided that the genetic and environmental factors are independent and that the misclassification of the environmental factor is independent of the genetic factor, both non-differential and differential misclassification of the environmental factor merely biases their interaction towards the null [208]. We found no meaningful association between the SNPs in *CYP17, ATM, CHEK2* and *ERBB2* and the environmental factors under study and thus conclude that recall bias is not a major problem in our study.

Misclassification
Misclassification or error in measuring the exposure or the outcome can be either differential or non-differential. Differential misclassification occurs when misclassification on one of the axes (exposure or outcome) is related to the other axis (exposure or outcome) [207]. Non-differential misclassification refers to errors on the one axis that are not related to the other axis [207]. For example, non-differential misclassification of the exposure occurs when the errors in the exposure classification are the same for cases and controls. Differential misclassification can bias the results either upwards or downwards but non-differential misclassification of dichotomous exposures causes bias towards the null. Non-differential misclassification of exposure variables with three or more categories is less predictable.

Misclassification of the Outcome
Misclassification of the outcome, such as the breast tumour characteristics, could have occurred in our study. The breast tumour characteristics were assessed by different pathologists and different laboratories throughout Sweden, which could have led to misclassification. However, the misclassification was most likely not related to the exposure as the pathologists or the laboratories had no knowledge of the individual's

genetic make-up. For example, estrogen and progesterone receptor status of the breast tumours and S-phase fraction were assessed at seven different laboratories across Sweden, but it is doubtful that genotype frequencies are related to inter-laboratory differences. Furthermore, a large proportion of the information on receptor status, S-phase fraction and grade for the breast tumours was missing. Assessment of receptor status and S-phase fraction was to a large extent dependent on the size of the tumour, but evaluation of grade was mostly dependent on the pathologist's decision. As genotype frequencies were not related to tumour size in our dataset, bias due to missing information on these factors seems unlikely.

Information from the Causes of Death Registry in Sweden has been found to be of high quality [209]. Thus, misclassification of a death as breast cancer death is unlikely. Furthermore, the same pathologist assessed the histological specimens for all of the endometrial cancer cases. This pathologist did not have any knowledge of the genetic make-up of the individuals who the samples belonged to and any misclassification that occurred is therefore merely non-differential.

Misclassification of the Exposure

Misclassification of the exposure in genetic association studies is related to the genotyping accuracy of the methods used and the quality control performed. Genotyping accuracy is essential in genetic association studies where the effect of interest is small. Using genotyping methods with low error rates is therefore crucial. Genotyping errors stem from various factors: a) Variation in DNA sequence (a mutation close to the marker site prevents amplification); b) low quantity or quality of DNA (only one of the two alleles present at a heterozygous locus is amplified or contaminant molecules are amplified as they have a higher probability of being amplified when the number of template DNA is low); c) biochemical artifacts (the *Taq* polymerase has a tendency to add a non-templated nucleotide to the 3´ end of the PCR product which creates an artificial band on the readout gel); d) human factor (sample mix-up, contamination, incorrect reagents added, pipetting error, data handling) [210].

Each 1% increase in non-differential genotyping error has been suggested to require a 2-8% increase in sample size in order to retain the same power in the study [211]. The Sequenom and Illumina methods used in our study are both highly automated methods – from sample handling to allele scoring – with low error rates (0.5% and 0.3%, respectively). Our power should therefore be similar to a study with 4% or 2.4% (depending on the method used) less sample size. The four percent corresponds to loosing 63 breast cancer cases, 28 endometrial cancer cases, and 63 controls, which only lowered our power by approximately 2% in general.

It is imperative in genetic association studies to randomly assign the case and control samples to the genotyping plates so any genotyping error does not become differential. Positive and negative controls should be added on each genotyping plate to assess contamination. If contamination occurs, the assay should be repeated. Furthermore, genotyping personnel should be blinded to case-control status in order to prevent any systematic differences between cases and controls, and concordance of genotypes should be assessed with different genotyping methods to validate genotype frequencies. We included positive and negative controls on all genotyping plates, we assigned the DNA samples randomly on the plates, our genotyping personnel were blinded to case-control status, and we replicated genotype calls with a separate genotyping method with over 99.5% concordance. Hence, differential misclassification of the exposure is unlikely to have accounted for our results.

<u>Loss to Follow-up</u>

Loss to follow-up refers to when subjects can no longer be located in follow-up studies. This is a potential problem since it can not be determined whether the losses are differential or non-differential because outcome information is unknown [207]. It is therefore important to maintain high and similar follow-up rates between the exposed and unexposed groups in follow-up studies. Since we relied on the constantly updated and nation-wide Causes of Death registry for determination of the outcome in the survival component of our study, loss to follow-up was mainly related to emigration of the women. However, only two women had emigrated in the parent breast cancer study. These two women were not selected for the current genetic study so loss to follow-up did not pose a problem in our study.

Length and Lead Time Bias

Length and lead time biases are not a problem in risk analyses but can be a problem in survival analyses.

Screening tends to detect cases with less aggressive forms of disease and who have longer survival. Length bias makes a screening program appear to be beneficial with regard to survival since people who are destined to have a favorable course are selectively identified [207]. If a person's genetic make-up determines whether this person will be screened or not, it would seem that the genetic factors under study were related to longer survival. Women using menopausal hormones are more likely than other women to undergo mammography screening. However, as the genotype frequencies were not related to method of breast cancer detection in our study, or to the use of menopausal hormones, length bias is most likely not a reason for concern in our study.

Lead time is the time from when a disease is detected by screening to when symptoms appear and the disease should have been detected [207]. Survival time of the cases therefore depends on whether the disease was detected by screening or whether the disease was detected because of symptoms. However, as mentioned above, the genotype frequencies in our study did not vary with the method of breast cancer detection or menopausal hormone use. Hence, we believe that problems related to lead time bias are negligible in our study.

Confounding

Until recently, a confounder has been defined according to three criteria: a) The exposure and confounder are associated, b) the confounder is associated with the outcome conditional on the exposure and c) the confounder is not an intermediate in the pathway between the exposure and the outcome.

Statistical association between two factors – such as the exposure and a confounder – occurs when one is the cause of the other, when they share a common cause, or both [206]. Hence, the criteria that the confounder can not be in the causal pathway between the exposure and the outcome (i.e. that exposure causes the confounder), implies that the confounder must cause the exposure or that they have a common cause (which would then in turn be the confounder) (Figure 5). The three criteria can therefore be redefined as a single criterion: Confounding is the presence of common causes to the exposure and outcome [206]. In this case, the confounder does not necessarily have to be a direct common cause of the exposure and outcome (Figure 5, b), but can also be a common cause indirectly through a surrogate marker for example (Figure 5, c).

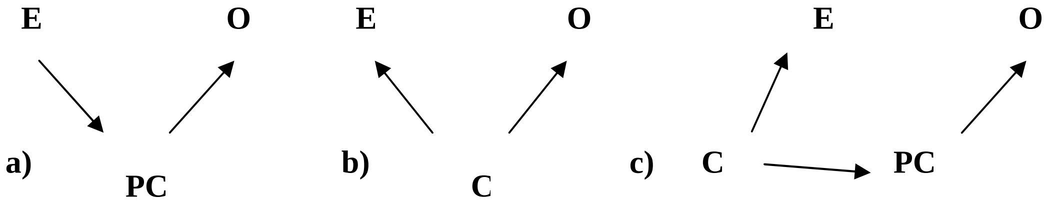

Figure 5. Examples showing the three possibilities of a statistical association between the exposure and a confounder. E=exposure, PC=potential confounder, O=outcome, C=confounder. a) E causes PC, which in turn causes O. Hence, PC is an intermediate in the pathway between E and O and not a confounder. b) C causes both E and O. C is therefore a confounder. c) C causes both E and PC, and PC causes O. C is thus the confounder, but PC can be adjusted for as the surrogate confounder.

Because a confounder should cause the exposure as well as the outcome either directly or indirectly – not merely be associated with the exposure and cause the outcome – confounding by any environmental factor is difficult to imagine in genetic association studies. No reproductive or lifestyle factor of an individual should be able to affect that person's genetic make-up. I believe that adjusting for potential confounders (apart from the matching factors) in genetic association studies when assessing the effect of the gene on the disease – either as main effects or stratified by environmental factors – should be avoided. Furthermore, applying statistical models using forward or backward selection of potential confounders to statistically assess the best fit of the different models should be discouraged. Instead, since our knowledge is limited regarding which environmental factors might be intermediates in the causal pathway between the genetic factors and disease; biological reasoning according to the above single criterion – and not the three criteria – should be applied to determine whether a factor is a confounder.

Unnecessary adjustments for potential confounders in genetic association studies can reduce the sample size considerably in the applied multiple regression models and can therefore reduce the power to a large extent. Power is an important issue in genetic association studies since the role of chance (due to lack of prior probability, see 'Random Error', 'Multiple Comparisons', 'Prior Probability' below) is one of the main challenges faced by this type of studies. For example, if we had adjusted for all the 'potential confounders' in the final models of our breast cancer study, our sample size would have been effectively reduced from 1579 breast cancer cases and 1516 controls to 1160 cases and 1106 controls in the regression analyses, or by 26.5% and 27%, respectively.

Confounding by Ethnicity

Confounding has been defined as mixing of effects between the causal factor and the confounder. Population stratification is a type of confounding and can be a problem in genetic association studies. It happens when the population under study is a mixture of two populations with different disease prevalence and different allele frequencies that are also different between cases and controls. For example, assume that a population being studied consists of Caucasians and Asians who differ in disease and exposure prevalence. The Caucasians tend to develop a certain type of cancer more often than the Asians and have a minor allele frequency (MAF) of 30% for a SNP under study, whilst the MAF is 10% among the Asians. The cases in this population will thus tend to be Caucasian with MAF of 30% whilst the controls will tend to be Asian with MAF of 10%. Hence, the SNP under study will erroneously be thought to increase the risk of the cancer.

There are several ways to adjust for population stratification in genetic association studies of unrelated individuals. First, the obvious procedure is to adjust for geographical region and markers of ethnic origin. This method has been stated to be sufficient to control for population stratification in populations with unrelated controls and no recent admixture [212]. However, other and more efficient methods have been suggested, such as genotyping anonymous genetic markers scattered throughout the genome that are independent of those affecting the disease of interest and that do not correlate with each other [213]. These markers should then reflect the baseline differences between cases and controls. They can be used to either estimate a scaling factor of the stratification to be incorporated into the association tests [214] or to subdivide the population into homogeneous subgroups [215]. However, Cardon and Palmer mention that in light of the limited empirical support for undetected population stratification as the major cause of false positive reports, it is currently not clear whether these extra genotyping efforts will be worthwhile in genetic association studies [216].

We believe that population stratification is of limited concern in our study since the participants were entirely Caucasian, Swedish born women residing in Sweden. All participants were born between 1919 and 1944, at a time when foreign immigration to Sweden was still rare [217]. It therefore seems likely that our population is relatively homogeneous with respect to genetic variation and that population stratification is not present in our study.

Random Error

Random error leads to a false association between the exposure and the outcome that arises from chance alone. This can result from an error in measuring the exposure or the outcome (for example human error occurring during documentation or calculation) or sampling variability [207]. Sampling is generally necessary in genetic association studies where expensive genetic tests are performed. Even if bias due to the selection of the participants does not exist, an unrepresentative sample of the source population can still be selected 'just by chance'. In the present genetic studies, we selected the participants from larger questionnaire-based studies. We first randomly chose an equal amount of cases and controls and then we over-sampled the long-term menopausal hormone users and the diabetics. Despite this over-sampling, Table 1 in Paper II and Table 1 in Paper IV show that the characteristics of the cases and controls were similarly distributed between the present and parent studies. Hence, our sub-samples were representative of the parent studies.

There are three principal ways to increase precision and reduce random error in epidemiological research [207]: (1) Increase the sample size of the study, (2) repeat the measurements within the study or repeat the entire study, and (3) design the study in such a way that the information obtained from a given sample size will be maximized. However, the absence of random errors does not guarantee the absence of systematic errors. It is possible to have precise but inaccurate findings due to bias for example.

Hypothesis Testing

Hypothesis testing is used to assess the role of random error in research [207]. First the null and alternative hypotheses are specified. The null hypothesis (H_O) states that there is no association between the exposure and disease whereas the alternative hypothesis (H_A) states that an association between the exposure and disease is present. Then a statistical test is performed to quantify the compatibility of the study data with the null hypothesis. The test statistic will yield a *P* value which is defined as the probability of obtaining the observed or more extreme result by chance alone, given that the null hypothesis is true. A *P* value of 0.02 indicates that there is only 2% probability of obtaining the observed result or one more extreme by chance alone if the null hypothesis is true. In this situation, chance is an unlikely explanation for the finding and we will reject the null hypothesis.

The decision whether a result can be called significant is made according to the level of significance we select. If we decide to reject the null hypothesis when the *P* value is less than 0.05, there is still 5% chance of rejecting the null hypothesis when it is in fact true. This is called the *type I* or *alpha error*. *Type II* or *beta error* occurs when the H_A is true but we fail to reject H_O.

Power

Power refers to the ability of a statistical test to correctly reject the null hypothesis when the alternative is true (1-*beta error*) [207]. When calculating power for a certain sample size, it is necessary to take into account a) the lowest magnitude of association that the study should be able to detect, b) the exposure prevalence in the control group, c) the prevalence of the disease in the population, and d) the selected level of significance. Power increases with increasing sample size, increasing magnitude of association or increasing prevalence of the exposure.

As mentioned above in the 'Background', under 'Linkage Disequilibrium', the relationship between a tagSNP and the SNP of interest in the gene is measured by the R^2 measure, which quantifies how well the tagSNP predicts the SNP of interest. The loss of power due to testing the SNP of interest indirectly is related to the R^2 measure [17, 19]. For example, assume the R^2 between a tagSNP and the SNP of interest in the gene is 0.8. This means that the tagSNP can predict the SNP of interest with 80% certainty, and that we would need to increase the sample size by 25% (1/0.8) in order to achieve the same power as if we would have tested the SNP of interest directly [17].

We wanted to quantify the power in our study, and in order to do so for the indirect analyses, we needed to take into account the ability of the tagSNPs to predict the other SNPs in the gene. Our study includes in theory three types of SNPs in each gene: The tagSNPs, (genotyped in all cases and controls), the observed SNPs (genotyped in 92 controls), and the unobserved SNPs (not genotyped and thus unknown). We could easily quantify how well the tagSNPs predicted the observed SNPs in our study and if we obtained high R^2 values we could easily have assumed that they are the same for the unobserved SNPs as well. However, this is not always a correct assumption [198]. We therefore assessed the capability of the tagSNPs to convey an association signal from unobserved SNPs as well as the observed SNPs (see 'Methods', 'Papers II-IV', 'Statistical Analyses', 'Coverage' above). We captured the unobserved SNPs with average R^2 of 0.92, 0.93 and 0.72 in *ATM*, *CHEK2* and *ERBB2*, respectively, and thus suffered minimal loss of power due to our indirect testing. Based on these R^2 values, we then calculated the power for risk assessment in both the breast cancer study and the endometrial cancer study. We calculated power related to assessment of survival in the breast cancer study using the Quanto program [204]. We were thus unable to take into account the R^2 value and calculated the power in accordance to direct analysis. The rare variants in *ATM* and *CHEK2* and the c.1-34T>C variant in *CYP17* were all tested directly in our studies and we calculated the power accordingly.

Breast Cancer Study

For the ability of haplotypes to predict the allele count at a causal locus with minor allele frequency of 0.20 – assuming $\alpha = 0.05$ – we had 89% power for *ATM*, 73% power for *ERBB2* and 87% power for *CHEK2* to detect an odds ratio of 1.3 in the risk component of the study.

To detect a hazard ratio of 1.4 with alpha level of 0.05 (assuming co-dominance) in the survival component of the study, we had 50% power for TAG1 in *CHEK2*, which had a minor allele frequency of 0.13, and 76% power for TAG5 in *CHEK2*, which had a minor allele frequency of 0.38.

Assuming α=0.05 and dominant inheritance, we had 68% power for the *CHEK2*1100delC* to detect a 2.25-fold increase in breast cancer risk. To detect an OR of 1.5, we had 43% power for the 2572 T→C in *ATM*, 53% power for the 4258 C→T in *ATM*, and 99.9% power for the *CYP17* c.1-34T>C variant.

Endometrial Cancer Study

For the ability of haplotypes to predict the allele count at a causal locus with minor allele frequency of 0.25 – assuming $\alpha = 0.05$ – we had 88% power for *ATM*, 88% power for *CHEK2* and 72% power for *ERBB2* to detect an odds ratio of 1.35.

Assuming α=0.05 and dominant inheritance to detect an OR of 1.5, we had 14% power for the *1100delC* in *CHEK2*, 33% power for the 2572 T→C in *ATM*, and 34% power for the 4258 C→T in *ATM*.

Multiple Comparisons

When performing one statistical test with an α-level of 0.05, there will be 5% probability of falsely rejecting the null hypothesis when it is true. However, when performing 10 statistical tests, the probability of observing at least one false positive finding will be 40% when the tests are independent $(1-(1-\alpha)^{k}$ where k applies to the number of tests performed). That is, the more tests one performs, the more likely it will be to eventually obtain a positive test. Thus, in order to retain the same overall rate of false positives (0.05), the significance level for each test has to become more stringent. The Bonferroni correction dictates that a new significance level of α/k should be applied to maintain the desired overall α-level. This procedure has on the other hand been criticised for being too conservative since the dependence of the statistical tests are not taken into account [218]. In our study, we corrected the global haplotype *P* values as well as individual SNP *P* values by applying the Westfall and Young permutation method for correction of the family-wise error rate, which takes into account the dependence structure of the hypotheses [202].

In Paper IV, we observed four statistically significant *P* values (<0.05) between the common tagSNPs or haplotypes and endometrial cancer risk: a) The overall *P* value for the *CHEK2* TAG1; b) the global *P* value for *CHEK2* haplotypes in non-smokers; and c) the global *P* values for *ATM* haplotypes in women who were younger than 49 years of age at time of menopause and in women with late age at last birth. None of these *P* values remained statistically significant after multiple testing correction. However, our *a priori* hypothesis stated that in interaction with increased estrogen exposure, these genes might affect endometrial cancer risk. We therefore assessed the effect of *CHEK2* TAG1 on endometrial cancer risk only among women with endometroid cancers – which are related to high estrogen exposure – and found a more pronounced association. In addition, when we statistically assessed the interaction between the haplotype that was responsible for each global haplotype association and the environmental factor in question, the interaction tests were highly statistically significant between *ATM* haplotype 4 and smoking status as well as *CHEK2* haplotype 1 and age at menopause. We therefore decided to report these three associations, discuss the biological mechanisms, and let the reader decide for him- or herself whether the associations will be worth attempting to replicate.

Prior Probability

It has been suggested that it is not the large number of tests in any one genetic association study that is the problem with multiple comparisons, but rather that each locus tested has such a small prior probability of being associated with the disease that even if the false positive rate is small, the vast majority of the positive findings will be false [12]. This might be the reason why the standards of statistical proof in classical epidemiology have become almost obsolete in genetic epidemiology. The problem is not necessarily the increased number of tests performed but rather the small prior belief that a certain SNP or a haplotype will be associated with the disease. Methods have been suggested to adjust for multiple testing by accounting for prior belief of an association [219], but they require knowledge of the prior probability of association. The very small prior probability that a SNP or a haplotype is associated with disease can instead be accounted for by applying a stringent significance threshold.

We detected a twofold increase in breast cancer risk related to the *1100delC* in *CHEK2*. Due to the functional effect of this deletion on the CHEK2 protein and the fact that a similar association with breast cancer risk had previously been reported in a large population-based study [32], we believe this deletion had much higher prior probability of being associated with breast cancer than the common polymorphisms in *CHEK2*. Hence, we report the association as a positive finding despite the borderline significance.

Effect Measure Modification

Effect measure modification means that an effect between exposure and outcome varies over strata of a third variable [207]. Its existence depends on the measure of association; i.e. when effect modification is present on the multiplicative scale, it will be absent on the additive scale and vice versa [207]. This fact has caused researchers some significant headaches. Rothman has stated that only an additive measure of effect modification can measure the underlying biological interaction of two factors [220]. However, calculating the appropriate measures of interaction for departure from additive risks is not straight forward in case-control studies since only surrogate measures – which are prone to bias – can be estimated [221]. Furthermore, additional covariates can not be accounted for in the models when assessing interaction using most of these surrogates measures [221]. For this reason, and because multiplicative models generally appear to be an adequate fit to observed data in practice [222], we assessed interaction on the multiplicative scale.

Paper I

The lack of association between *CYP17* c.1-34T>C and overall breast cancer risk in our data is in line with results from 14 previous studies – where 10 included only Caucasian women – and a recent meta-analysis [99, 103-116]. Two groups have reported an association between *CYP17* genotype and breast cancer risk in postmenopausal women [100, 101] and in contrast to Feigelson and colleagues, we did not find any association between this polymorphism and advanced breast cancer [102]. All three studies were performed in non-Caucasian populations.

A possible mechanism for the *CYP17* c.1-34T>C polymorphism to influence breast cancer risk is through increased biosynthesis of and therefore increased levels of circulating estrogen. Contrary to the predicted effect of the A2 allele, one group found decreased *CYP17* mRNA levels in A2 carriers [223]. Studies regarding association of *CYP17* c.1-34T>C and circulating hormone levels as well as markers of hormonal status (i.e. age at menarche or menopausal hormone use) have recently been reviewed [224].

Increased estradiol levels have been associated with the A2 allele in premenopausal women [100, 192] as well as in postmenopausal women [103, 225], but three groups did not report any significant changes in hormone levels by genotype; two in postmenopausal women [111, 226] and one in premenopausal women [227]. Results from seven [228-230] studies have indicated a moderate association between the A2 allele and earlier menarche; an association that was not detected in our study, nor in five others [103, 106, 115, 231, 232]. Furthermore, previous investigators have posited that *CYP17* genotype may be associated with use of menopausal hormones, an important risk factor for breast cancer, but results have been inconsistent [105, 106, 115, 233-235]. We found no such association in our data.

We identified interaction with age at menarche, but considered it unlikely that the A2/A2 genotype would increase breast cancer risk in women with age at menarche less than 13 years without also moderately increasing risk in women with age at menarche between 12 and 14 years. Instead, the A2/A2 genotype decreased risk in the latter group, which is difficult to explain biologically. A similar pattern was seen for the interaction with age at menopause. To our knowledge, other groups have not reported comparable findings and we therefore believe the results to be caused by chance alone.

Papers II-IV

Breast Cancer Risk

<u>tagSNPs and Haplotypes</u>

We found no association between common variation in the *CHEK2* gene and overall breast cancer risk, which was in agreement with earlier findings of common polymorphisms in *CHEK2* [162]. We correspondingly found no effect of common variation in *ATM* or *ERBB2* on breast cancer risk, even when stratified by known breast cancer risk factors. One study of *ATM* [146] and two studies of *ERBB2* [178, 181] are in agreement with our findings. Tamimi et al. found no association between haplotypes of five Hapmap tagSNPs (one of which was our TAG7) in *ATM* and breast cancer risk [146]. Benusiglio et al. explored *ERBB2* haplotypes composed of five tagSNPs (three of which were our TAG2, 3 and 5) – including the non-synonymous I655V and P1170A – in relation to breast cancer risk [178], whilst Han et al. solely studied the I655V and P1170A as tagSNPs [181]. Neither study found any effect of the haplotypes on breast cancer risk. Common haplotypes in *ERBB2* thus do not appear to affect breast cancer risk, although results regarding the I655V common variant in *ERBB2* have been conflicting [178-181]. We included the I655V as a tagSNP in our study and genotyped the P1170A in the 92 controls. We found no association of the I655V with breast cancer risk.

Three groups have found an association between specific *ATM* common haplotypes and breast cancer risk [141, 144, 145]. Lee et al. [141] and Koren et al. [145] reconstructed haplotypes in *ATM* from five and eight randomly selected common SNPs respectively, whereas Angele et al. [144] included eleven common SNPs in *ATM* for their haplotypes estimation that had either been previously reported in the literature or that they had detected by sequencing. SNP selection overlapped somewhat between the three studies, but none of them reported the likelihood of their SNPs being able to predict underlying variation in the gene. Furthermore, findings from two of the three groups [144, 145] were based on small sample sizes.

Rare Variants

We found more than twofold increase in breast cancer risk related to carriers of the *1100delC* in *CHEK2* compared to non-carriers, which is consistent with results from other Northern European populations [32, 236]. The *1100delC* variant has not previously been studied in the Swedish population. The variant is rare [157, 158, 161, 237-241], except in a few Northern European populations such as Finland and the Netherlands where moderate frequency of 1% or above has been observed [151, 152, 236]. The low population frequency of *1100delC* in the Swedish population is therefore generally in line with previous data.

In line with two [134, 144] out of three reports [134, 142, 144], we found an elevated breast cancer risk – though not significant – for carriers of the rare 4258 T allele in the *ATM* gene. We did not, on the other hand, find an association for carriers of the 2527 C allele. One study found twofold increase in breast cancer risk related to the 2527 T→C in a US population, but did not confirm the finding in a Polish population [143]. Three other groups did not detect any significant effect of the 2527 T→C on breast cancer risk [134, 142, 144], although one of the groups found elevated point estimates [134].

Breast Cancer Survival and Tumour Characteristics

Our data did not support an association between common variation in *ATM*, *CHEK2* and *ERBB2* with breast cancer survival or the risk of developing tumours of different characteristics. Hence, we did not confirm the finding of Han et al. where they found an *ERBB2* haplotype composed of two non-synonymous tagSNPs – I655V and P1170A – to increase the risk of breast cancer death or recurrence [181]. We found no effect of the I655V on breast cancer survival or tumour characteristics-defined breast cancer. To our knowledge, no study has investigated *ATM* or *CHEK2* common haplotypes in relation to breast cancer survival or tumour characteristics, although one study explored the effect of three common polymorphisms in *ATM* and two common polymorphisms in *CHEK2* on breast cancer survival [242]. They found no association, which is in agreement with our findings. The rare *1100delC* mutation in *CHEK2* has been associated with breast tumours of high grade [153, 154] as well as steroid receptor positive breast tumours, but not with overall survival [153]. The mutation was too rare in our population to be studied in relation to breast cancer survival or tumour characteristics.

Endometrial Cancer Risk

Until now, germ-line variation in the *ATM, CHEK2* and *ERBB2* genes has not previously been assessed in association with endometrial cancer risk.

CHEK2 TAG1

We found homozygous carriers of the minor allele of the common tagSNP TAG1 in *CHEK2* to be at increased risk of endometrial cancer. The rare allele of TAG1 was the only rare allele carried by haplotype 4 in *CHEK2* and we consequently found an increased risk for haplotype 4 carriers in *CHEK2*. The effect of TAG1 in *CHEK2* on endometrial cancer risk was stronger among endometroid tumours. Endometrial cancers can be divided into Type I endometroid tumours and Type II non-endometroid tumours [81-83], where endometroid tumours constitute the majority of endometrial cancers. The endometroid tumours appear to be the tumours that are mainly caused by estrogen exposure [81-83]. Estrogen metabolites have been reported to cause a number of DNA lesions both directly and indirectly through redox cycling processes [185]. Indirect damage includes single strand DNA breaks, 8-hydroxylation of guanine bases, and DNA adducts [185], whilst direct DNA damage caused by covalent binding of quinone intermediates of 4-hydroxyestrogens to DNA can result in the formation of mutagenic apurinic sites [243]. The estrogen metabolites 2- and 4-hydroxyestrogens have also been reported to cause double strand breaks *in vitro* [186]. DNA double strand breaks seem to be the predominant signal for the activation of ATM-mediated pathways [187]. The CHEK2 protein is activated by ATM and thus affects cell cycle arrest and DNA repair [118-121]. Our results imply that a defect in the *CHEK2* gene affecting the function or expression of the CHEK2 protein increases endometrial cancer risk mainly in combination with increased estrogen exposure. This study was designed in such a way that the tagSNPs in each gene predicted common variation of over 3% in minor allele frequency with at least 80% probability. It is unlikely that TAG1 itself has a structural effect on the CHEK2 protein as it is located in an intronic region, but it is still possible it has a regulatory effect on the protein expression. Another likely scenario is that a common polymorphism in linkage disequilibrium with TAG1 might be responsible for this association.

ATM Haplotype 4

Interestingly, we observed carriers of haplotype 4 in *ATM* to have decreased endometrial cancer risk if they had never smoked in their lifetime. Carriers of this haplotype also had decreased endometrial cancer risk overall, although it was not as pronounced as in non-smokers and did not carry over to the global test of significance. Haplotype 4 did not carry a rare allele from any of the tagSNPs (it carried only the tagSNP common alleles), which is in line with the observed lack of effect of the *ATM* tagSNPs on endometrial cancer risk. One plausible biological explanation for this finding is that non-smoking *ATM* haplotype 4 carriers are more efficient in repairing estrogen-related DNA damage than non-carriers. Smoking has been suggested to have anti-estrogenic effects [244] and women who smoke therefore are likely to be less exposed to estrogen. These women may be able to adequately repair the lower levels of estrogen-related DNA damage regardless of their *ATM* haplotype. In non-smokers however, estrogen levels have been found to be higher than in smokers [245-247]. In this situation, the increased DNA damage may exceed the repair-capabilities of those women who do not possess *ATM* haplotype 4, whereas women with *ATM* haplotype 4 may be able to manage the excess levels of damage imposed by estrogen.

CHEK2 Haplotype 1

Additionally, we found decreased endometrial cancer risk among carriers of haplotype 1 in *CHEK2* who were younger than 49 years of age at time of menopause. The only rare allele carried by haplotype 1 was the C allele of TAG5 in *CHEK2*, but TAG5 itself did not appear to affect endometrial cancer risk. Women who are relatively young at time of menopause have experienced fewer ovulatory cycles and thus less exposure to estrogen than women who experience menopause at an older age [248]. It is possible that carriers of haplotype 1 are more capable of managing the low estrogen-related DNA damage in women with early age at menopause than carriers of other haplotypes in *CHEK2*. However, in the presence of high estrogen-related DNA damage (in women with higher estrogen exposure), even haplotype 1 carriers appear to be unable to repair the large amount of DNA damage.

Rare Variants

Neither the *1100delC* deletion in *CHEK2* nor the rs1800056 (2572 T→C, F858L) and rs1800058 (4258 C→T, L1420F) variants in *ATM* have been previously studied in relation to endometrial cancer risk. We found no effect of these variants on endometrial cancer risk in our data, but since there were very few carriers of the rare alleles, our statistical power was low.

Causation

In this thesis, I report associations between the rare *1100delC* in *CHEK2* and risk of breast cancer as well as between carriers of common variation in *CHEK2* and *ATM* and endometrial cancer risk.

I have assessed the role of bias and confounding and come to the conclusion that the findings are unlikely to be due to these factors. I could however not exclude the role of random error in the endometrial cancer findings, but decided nevertheless to consider them plausible until refuted.

Austin Bradford Hill suggested a set of nine guidelines for researchers to aid them in assessing if associations are causal [249]. None of the guidelines were intended to be rigid criteria, but rather as imperfect guides towards causation. Below I discuss each guideline in relation to our findings:

1. ***Strength of the association***
   Strong associations are more likely to be causal as they are unlikely to be accounted for entirely by, for example, bias and confounding.
   All four associations were relatively strong and therefore any bias or confounding by ethnicity that might have occurred is unlikely to have entirely accounted for our results.

2. ***Consistency***
   Associations are more likely to be causal if they are observed repeatedly by different researchers.
   Previous studies had found a similar association between the *1100delC* and breast cancer risk [32], but the endometrial cancer findings await replication.

3. ***Specificity***
   The concept of specificity means that a cause should only lead to a single effect and vice versa.
   There are numerous well known exceptions to this concept and epidemiologists therefore do not consider specificity a useful guideline.

4. ***Temporality***
   The belief that a cause must precede the disease is a well known and accepted requirement for causality.
   The exposure in our studies obviously preceded the disease as the genetic make-up of the women was determined at conception.

5. ***Biological Gradient***
   An association is more likely to be causal if its strength increases as the exposure level increases.
   The effect of the *1100delC* and TAG1 in *CHEK2* increased from heterozygous to homozygous carriers of the rare allele. The effect estimates for the haplotype effects referred to heterozygous carriers of the haplotypes and should be squared to obtain the estimates for homozygous carriers the haplotypes.

6. ***Plausibility***
   There should be existing biological or social model to explain the association.
   The *ATM* and *CHEK2* genes respond to DNA damage caused by estrogen, and increased estrogen exposure can lead to the development of breast or endometrial cancer.

7. ***Coherence***
   The interpretation of the data should not seriously conflict with generally known facts of the natural history and biology of the disease.
   Both breast and endometrial cancer are believed to be in part due to genetic factors. Thus, finding genes that affect the risk of these diseases is highly probable.

8. ***Experiment***
   Intervention that modifies the exposure through prevention, treatment or removal should result in less disease.
   Intervention has not yet become feasible in genetic epidemiology since gene therapy is still in its early stages and the numerous ethical issues involved have not been addressed.

9. ***Analogy***
   Analogies should exist between the observed association and other associations.
   Other genes with similar functions as *ATM* and *CHEK2* could also be involved in the development of breast or endometrial cancers. For example, the function of the *BRCA1* gene is highly linked to the *ATM* and *CHEK2* gene functions [250,

251] and mutations in the *BRCA1* gene have been shown to be involved in breast cancer development [34].

Implications and Future Research

Replication is essential in genetic association studies as well as other epidemiological studies. We want to make sure that the observed association is not due to bias, other factors or chance. The best way to ensure this is by testing the same association in completely independent populations, preferably by different investigators. Whilst the benefits with respect to public health may not be immediate, confirmation of the results in independent populations is still worthwhile. It will not turn out to be beneficial to individually correct the genetic variation affecting the disease aetiology as the prevalence of each variation will be high in the population, whilst the penetrance will be low. In contrast, the important role of genetic association studies will be to elucidate the biological mechanisms that lead to disease. This elucidation will specifically be of importance in order to identify susceptible subgroups of people in which intervention of certain environmental factors will be especially beneficial.

In light of the positive findings in this thesis, what are the obvious subsequent steps? The first step would be to replicate the endometrial cancer findings in another larger and independent population. Secondly, it would be of interest to sequence the carriers of the *CHEK2* TAG1 rare allele as well as the women carrying haplotype 1 in *CHEK2* and haplotype 4 in *ATM*. Comparison of the sequence information of these women with non-carriers of TAG1 and the respective haplotypes could lead to detection of the causal variants responsible for the observed associations.

It would also be of interest to genotype the *CHEK2*1100delC* in a much larger population in order to evaluate gene-environment interaction and assess the effect of the deletion on breast cancer survival or tumour characteristics.

CONCLUSIONS

Paper I

- Genetic variation in *CYP17* had no obvious effect on breast cancer risk regardless of histopathology or menopausal hormone use.

Papers II-IV

Common Variation

- Common variation in the *ATM*, *CHEK2* or *ERBB2* genes did not affect breast cancer risk overall or in combination with breast cancer risk factors.
- Common variation in the *ATM*, *CHEK2* or *ERBB2* genes was not associated with the risk of tumour characteristics-defined breast cancer or breast cancer death.
- Common variation in the *ERBB2* gene did not show any relationship with the risk of endometrial cancer and the *ATM* and *CHEK2* genes did not appear to affect overall endometrial cancer risk.
- Homozygous carriers of the rare allele of TAG1 in *CHEK2* had more than twice the risk of developing endometroid endometrial cancer, compared to non-carriers.
- Non-smoking carriers of haplotype 4 in *ATM* possessed half the endometrial cancer risk of non-carriers.
- Among carriers of haplotype 1 in *CHEK2* who had experienced menopause below 49 years, endometrial cancer risk was halved compared to non-carriers.

Rare Variants

- Carriers of the rare *1100delC* deletion in *CHEK2* had a more than twofold increased breast cancer risk compared to non-carriers, but the deletion did not seem to have an effect on endometrial cancer risk.
- Rare variants in the *ATM* gene did not appear to affect breast or endometrial cancer risk.

ÁGRIP Á ÍSLENSKU

Brjóstakrabbamein er algengasta krabbameinið meðal kvenna í heiminum í dag og krabbamein í legi er algengasta krabbameinið í æxlunarfærum kvenna í hinum iðnvædda heimi. Konur með fjölskyldusögu um brjósta- eða legkrabbamein eru í tvöfalt meiri hættu á að mynda þessi krabbamein en aðrar konur. Erfðafræðilegir áhættuþættir legkrabbameina almennt og brjóstakrabbameina í konum, sem ekki bera sjaldgæfar stökkbreytingar með háa sýnd, eru að stórum hluta óþekktir. Því hefur verið sett fram sú tilgáta að fjölgena líkan skýri eftirstöðvarnar af arfgengi krabbameinanna. Þetta líkan gerir ráð fyrir samspili margra algengra genabreytileika með litla sýnd sem eru auk þess taldir hafa áhrif í samspili við hina ýmsu umhverfisþætti. Við ákváðum þar af leiðandi að rannsaka áhrif algengra breytileika í lykil brjósta- og legkrabbameinsgenum á almenna brjósta- og legkrabbameinsáhættu. Við rannsökuðum einnig hvort að breytileikarnir hafi áhrif í samspili við umhverfisþætti, hvort þeir auki hættu á brjóstakrabbameinum með ákveðin einkenni, eða hvort þeir auki hættu á að brjóstakrabbamein leiði til dauða. Við rannsökuðum 1579 konur með brjóstakrabbamein, 705 konur með legkrabbamein og 1565 heilbrigð viðmið. Allir þáttakendur gáfu vefja- og blóðsýni og skiluðu inn spurningalistum með upplýsingum um hina ýmsu lífshætti.

Talið er að *CYP17*, *ATM*, *CHEK2* og *ERBB2* genin leiki hlutverk í myndun og þróun krabbameina. Hlutverk þessara gena í myndun brjósta- og legkrabbameina liggja í áhrifum þeirra á efnaskipti estrógena, virkjun DNA viðgerðakerfa og fjölgun fruma. Í rannsókn okkar greindum við í öllum þáttakendunum arfgerð algengra breytileika og sjaldgæfra stökkbreytinga í þessum genum. Við mátum síðan tengsl þessara breytileika og haplótýpa þeirra við krabbameinsáhættu og lifun.

Í rannsókn okkar kom í ljós að arfberar hinnar sjaldgæfu *1100delC* stökkbreytingar í *CHEK2* voru algengari meðal kvenna með brjóstakrabbamein en heilbrigðra viðmiða. Arfberarnir voru í tvöfalt meiri hættu á að mynda brjóstakrabbamein samanborið við stofngerðir (LH 2.26, 95% ÖM 0.99-5.15). Niðurstöður okkar bentu einnig til þess að arfhreinir arfberar rs4987886 breytileikans í *CHEK2* hafi aukna hættu á legkrabbameini af gerð I (P = 0.005) samanborið við stofngerðir. Vid fundum auk þess verndandi áhrif gegn legkrabbameini meðal arfbera haplótýpu í *ATM* sem ekki reykja (P = 0.0007) og meðal arfbera haplótýpu í *CHEK2* sem voru yngri en 49 ára við tíðahvörf (P = 0.0009), samanborið við arfbera annarra haplótýpa. *ATM, CHEK2* og *ERBB2* genin virtust ekki leika hlutverk í myndun brjóstakrabbameina með ákveðin einkenni eða í lifun kvenna með brjóstakrabbamein. Við fundum engin tengsl milli *CYP17, ATM* og *ERBB2* genanna og brjóstakrabbameinsáhættu. *ERBB2* genið virtist auk þess ekki hafa áhrif á myndun legkrabbameins.

Mat okkar á brjóstakrabbameinsáhættunni tengdri *CHEK2*1100delC* stökkbreytingunni samsvarar niðurstöðum annarra rannsókna meðal Norður-Evrópuþjóða. Frekari rannsókna er hins vegar þörf í sambandi vid tengsl *CHEK2* og *ATM* við legkrabbameinsáhættu þar sem niðurstöður okkar voru ekki tölfræðilega marktækar þegar við höfðum tekið til greina þann fjölda tölfræðiprófa sem voru framkvæmd.

ACKNOWLEDGEMENTS

During the four years of my PhD studies, I spent time both in Sweden and Singapore and many people contributed either directly or indirectly to my work. They all deserve my gratitude. Specifically, I would like to thank:

Sara Wedrén, my main supervisor, for your dedication and indispensable guidance. For always encouraging me, for your support, your compliments and your honesty. Your knowledge in the field continues to inspire me. It has been an extremely valuable experience working with you.

Keith Humphreys, my co-supervisor, for guiding me through the jungle of statistics and for teaching me the statistical necessities of association studies. Without your work, these papers would never have been complete.

Per Hall, my co-supervisor, for creating an enjoyable working environment at level 5, for interesting discussions and for always being willing to help.

Hans-Olov Adami, former chair of MEB, for inviting me to visit MEB for an interview after receiving my job application and for introducing me to my supervisors to be.

Jianjun Liu, my co-supervisor, for your guidance during my time at the Genome Institute of Singapore (GIS). For always sharing your quick mind and your vast genetic and biological knowledge.

Kee Seng Chia, my co-supervisor, and **Edison Liu**, director of GIS, for valuable collaboration.

Carine Bonnard, my colleague at GIS, for being my best friend in Singapore. For teaching me all the wonders of excel data mining and for your guidance through my genotyping work. For fun lunches, coffees, dinners and parties. For going out jogging with me in the melting Singapore heat and for fun Saturdays by the pool. **Carine** and **Olivier**, for extremely enjoyable and humorous couple dinners and tennis/running evenings with Anthony and me.

Sonja Davila, a GIS colleague, for fun lunches, coffees and dinners. **Yuqing Li, Yi Li** and **other colleagues at GIS** for interesting collaboration, for enjoyable working environment and for introducing me to the Singaporean food courts in Holland Village.

Joseph Abraham, for being the most enthusiastic teacher I have ever had, and for always being ready to answer epidemiological questions, even across the Atlantic Ocean.

Lena Rosenberg, fellow PhD student and co-author, for always very inspiring conversations and for all your help with the dissertation process and planning.

Stephanie Bakaysa, for being such a good friend at MEB, for our sun tanning times and for all the good times going out partying.

Arvid Sjölander, **Guðrún Jónasdóttir**, and **Paul Dickman**, for all your help with statistics and epidemiology. **Guðrún** for being such a good friend.

Katarina Olofsson, **Anna Svensson**, **Åsa Odenbro** and **Chantal Mascarenhas,** for all the great times during the GAME course in Singapore.

Linda Lindström, **Unnur Valdimarsdóttir, Ulrika Eriksson, Lorenzo Richiardi, Pär Sparén, all members of Pelles group,** and all **other MEBers** for many good times and for creating a wonderful working environment here at MEB. **Linda** for all the good advice in the last few months.

My friends in Stockholm: **Camilla Oxhamre, Keira Melican, Paula Maguire, Guðmundur Bergsson, Mikael Stabo, Bartek Zube, Mark Iles, Jenny Barton, Mark Wareing** and **Bradley Goodfellow** for always being willing to share food and/or a beer and for great MTC parties. **Camilla** and **Keira** for being wonderful friends, for all your support, for the best fikas and the best girls' night out.

My Icelandic friends: **Jóhannes Elmar Jóhannesson** and **Linda Björk Ólafsdóttir** for always making my visits back home so enjoyable with long conversations lasting late in the night. **Lára Björg Lárusdóttir,** for being my best friend in Junior College and for still being a good friend even though we are far apart. **Þórgunnur E. Pétursdóttir** for our great times together in Stockholm and later in Iceland. For all the cinema trips, going out partying and the fun gossip. **Linda** and **Þórgunnur** for organizing the ultimate hens night.

Christopher Gunnell, **Jennifer Gunnell and the kids** for letting us stay with you for 5 weeks during our visit to Perth, for the numerous and fantastic Xmas and birthday pressies, and for just being a great family. **Chris** for always being willing to answer my questions.

John Gunnell and **Glenys Gunnell,** my parents-in-law, for your love and support.

My sisters; **Árnína Björg Einarsdóttir** and **Guðný Dóra Einarsdóttir,** for always being there with your advice and support. You are the best sisters a girl could have!

My parents; **Einar Guðni Njálsson** and **Sigurbjörg Bjarnadóttir,** for your support in all aspects of life. For always encouraging me to seek education. My mother for always reminding me that education is the only way for women to become independent.

My husband and best friend; **Anthony Stephen Gunnell.** Well, what can I say, we came to Sweden together, we work at the same department, our offices are in the same corridor, and we frequently visit each other offices with questions about work related issues. We share our life experiences, both at home and at work. Words cannot describe my gratitude. I just hope we will always be together and that our long conversations about everything will not fade with the years.

REFERENCE LIST

1. Strachan T and Read AP. **Human Molecular Genetics 2**, 1999. Oxford, UK: BIOS Scientific Publishers Ltd.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, et al. **The sequence of the human genome**. *Science*, 2001. 291(5507):1304-1351.
3. Burton PR, Tobin MD, and Hopper JL. **Key concepts in genetic epidemiology**. *Lancet*, 2005. 366(9489):941-951.
4. Schork NJ, Fallin D, and Lanchbury JS. **Single nucleotide polymorphisms and the future of genetic epidemiology**. *Clin Genet*, 2000. 58(4):250-264.
5. Gray IC, Campbell DA, and Spurr NK. **Single nucleotide polymorphisms as tools in human genetics**. *Hum Mol Genet*, 2000. 9(16):2403-2408.
6. Nelen MR, Padberg GW, Peeters EA, Lin AY, van den Helm B, Frants RR, Coulon V, Goldstein AM, van Reen MM, Easton DF, Eeles RA, Hodgsen S, Mulvihill JJ, Murday VA, Tucker MA, et al. **Localization of the gene for Cowden disease to chromosome 10q22-23**. *Nat Genet*, 1996. 13(1):114-116.
7. Buchwald M, Zsiga M, Markiewicz D, Plavsic N, Kennedy D, Zengerling S, Willard HF, Tsipouras P, Schmiegelow K, Schwartz M, and et al. **Linkage of cystic fibrosis to the pro alpha 2(I) collagen gene, COL1A2, on chromosome 7**. *Cytogenet Cell Genet*, 1986. 41(4):234-239.
8. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, and King MC. **Linkage of early-onset familial breast cancer to chromosome 17q21**. *Science*, 1990. 250(4988):1684-1689.
9. Kruglyak L and Nickerson DA. **Variation is the spice of life**. *Nat Genet*, 2001. 27(3):234-236.
10. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, et al. **Characterization of single-nucleotide polymorphisms in coding regions of human genes**. *Nat Genet*, 1999. 22(3):231-238.
11. Brookes AJ. **The essence of SNPs**. *Gene*, 1999. 234(2):177-186.
12. Cordell HJ and Clayton DG. **Genetic association studies**. *Lancet*, 2005. 366(9491):1121-1131.
13. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, and Todd JA. **A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region**. *Nat Genet*, 2006. 38(6):617-619.
14. Ardlie KG, Kruglyak L, and Seielstad M. **Patterns of linkage disequilibrium in the human genome**. *Nat Rev Genet*, 2002. 3(4):299-309.
15. Palmer LJ and Cardon LR. **Shaking the tree: mapping complex disease genes with linkage disequilibrium**. *Lancet*, 2005. 366(9492):1223-1234.
16. Wall JD and Pritchard JK. **Haplotype blocks and linkage disequilibrium in the human genome**. *Nat Rev Genet*, 2003. 4(8):587-597.

17. Pritchard JK and Przeworski M. **Linkage disequilibrium in humans: models and data**. *Am J Hum Genet*, 2001. 69(1):1-14.

18. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, and Lander ES. **High-resolution haplotype structure in the human genome**. *Nat Genet*, 2001. 29(2):229-232.

19. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, et al. **The structure of haplotype blocks in the human genome**. *Science*, 2002. 296(5576):2225-2229.

20. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, et al. **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21**. *Science*, 2001. 294(5547):1719-1723.

21. Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, et al. **Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots**. *Nat Genet*, 2003. 33(3):382-387.

22. Stenzel A, Lu T, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, and Schreiber S. **Patterns of linkage disequilibrium in the MHC region on human chromosome 6p**. *Hum Genet*, 2004. 114(4):377-385.

23. Goldstein DB, Ahmadi KR, Weale ME, and Wood NW. **Genome scans and candidate gene approaches in the study of common diseases and variable drug responses**. *Trends Genet*, 2003. 19(11):615-622.

24. Cardon LR and Abecasis GR. **Using haplotype blocks to map human complex trait loci**. *Trends Genet*, 2003. 19(3):135-140.

25. Chakravarti A. **Population genetics--making sense out of sequence**. *Nat Genet*, 1999. 21(1 Suppl):56-60.

26. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, and Pericak-Vance MA. **Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families**. *Science*, 1993. 261(5123):921-923.

27. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, and Reitsma PH. **Mutation in blood coagulation factor V associated with resistance to activated protein C**. *Nature*, 1994. 369(6475):64-67.

28. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, et al. **The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes**. *Nat Genet*, 2000. 26(1):76-80.

29. Smith DJ and Lusis AJ. **The allelic structure of common disease**. *Hum Mol Genet*, 2002. 11(20):2455-2461.

30. Wang WY, Barratt BJ, Clayton DG, and Todd JA. **Genome-wide association studies: theoretical and practical concerns**. *Nat Rev Genet*, 2005. 6(2):109-118.

31. Lohmueller KE, Pearce CL, Pike M, Lander ES, and Hirschhorn JN. **Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease**. *Nat Genet*, 2003. 33(2):177-182.

32. **CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies**. *Am J Hum Genet*, 2004. 74(6):1175-1182.

33. Antoniou AC, Pharoah PD, McMullan G, Day NE, Stratton MR, Peto J, Ponder BJ, and Easton DF. **A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes**. *Br J Cancer*, 2002. 86(1):76-83.

34. **Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group**. *Br J Cancer*, 2000. 83(10):1301-1308.

35. Pagani F and Baralle FE. **Genomic variants in exons and introns: identifying the splicing spoilers**. *Nat Rev Genet*, 2004. 5(5):389-396.

36. Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, Buckland PR, and O'Donovan MC. **Functional analysis of human promoter polymorphisms**. *Hum Mol Genet*, 2003. 12(18):2249-2254.

37. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, and Lee MP. **Allelic variation in gene expression is common in the human genome**. *Genome Res*, 2003. 13(8):1855-1862.

38. Parkin DM, Bray F, Ferlay J, and Pisani P. **Estimating the world cancer burden: Globocan 2000**. *Int J Cancer*, 2001. 94(2):153-156.

39. **Cancer Incidence in Sweden 2004**, 2006: The National Board of Health and Welfare, Centre for Epidemiology, Official Statistics of Sweden.

40. **Canceröverlevnad i Sverige 1960-1998 – utvecklingen över fyra decennier**, 2003: The National Board of Health and Welfare, Centre for Epidemiology, Official Statistics of Sweden.

41. Bray F, McCarron P, and Parkin DM. **The changing global patterns of female breast cancer incidence and mortality**. *Breast Cancer Res*, 2004. 6(6):229-239.

42. Chia KS, Reilly M, Tan CS, Lee J, Pawitan Y, Adami HO, Hall P, and Mow B. **Profound changes in breast cancer incidence may reflect changes into a Westernized lifestyle: a comparative population-based study in Singapore and Sweden**. *Int J Cancer*, 2005. 113(2):302-306.

43. Magnusson CM, Persson IR, Baron JA, Ekbom A, Bergstrom R, and Adami HO. **The role of reproductive factors and use of oral contraceptives in the aetiology of breast cancer in women aged 50 to 74 years**. *Int J Cancer*, 1999. 80(2):231-236.

44. Beral V. **Breast cancer and hormone-replacement therapy in the Million Women Study**. *Lancet*, 2003. 362(9382):419-427.

45. Magnusson C, Baron JA, Correia N, Bergstrom R, Adami HO, and Persson I. **Breast-cancer risk following long-term oestrogen- and oestrogen-progestin-replacement therapy**. *Int J Cancer*, 1999. 81(3):339-344.

46. **Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women**

without breast cancer from 54 epidemiological studies. Collaborative Group on Hormonal Factors in Breast Cancer. *Lancet*, 1996. 347(9017):1713-1727.

47. Friedenreich CM. **Review of anthropometric factors and breast cancer risk**. *Eur J Cancer Prev*, 2001. 10(1):15-32.

48. Feigelson HS, Jonas CR, Teras LR, Thun MJ, and Calle EE. **Weight gain, body mass index, hormone replacement therapy, and postmenopausal breast cancer in a large prospective study**. *Cancer Epidemiol Biomarkers Prev*, 2004. 13(2):220-224.

49. Hamajima N, Hirose K, Tajima K, Rohan T, Calle EE, Heath CW, Jr., Coates RJ, Liff JM, Talamini R, Chantarakul N, Koetsawang S, Rachawat D, Morabia A, Schuman L, Stewart W, et al. **Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease**. *Br J Cancer*, 2002. 87(11):1234-1245.

50. Vainio H, Kaaks R, and Bianchini F. **Weight control and physical activity in cancer prevention: international evaluation of the evidence**. *Eur J Cancer Prev*, 2002. 11 Suppl 2:S94-100.

51. Lagerros YT, Hsieh SF, and Hsieh CC. **Physical activity in adolescence and young adulthood and breast cancer risk: a quantitative review**. *Eur J Cancer Prev*, 2004. 13(1):5-12.

52. **Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease**. *Lancet*, 2002. 360(9328):187-195.

53. Russo J, Moral R, Balogh GA, Mailo D, and Russo IH. **The protective role of pregnancy in breast cancer**. *Breast Cancer Res*, 2005. 7(3):131-142.

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

55. Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, Vierkant RA, Maloney SD, Pankratz VS, Hillman DW, Suman VJ, Johnson J, Blake C, Tlsty T, Vachon CM, et al. **Benign breast disease and the risk of breast cancer**. *N Engl J Med*, 2005. 353(3):229-237.

56. Wolf I, Sadetzki S, Catane R, Karasik A, and Kaufman B. **Diabetes mellitus and breast cancer**. *Lancet Oncol*, 2005. 6(2):103-111.

57. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, and Paterson AD. **Mammographic breast density as an intermediate phenotype for breast cancer**. *Lancet Oncol*, 2005. 6(10):798-808.

58. Pharoah PD, Day NE, Duffy S, Easton DF, and Ponder BA. **Family history and the risk of breast cancer: a systematic review and meta-analysis**. *Int J Cancer*, 1997. 71(5):800-809.

59. Magnusson C, Colditz G, Rosner B, Bergstrom R, and Persson I. **Association of family history and other risk factors with breast cancer risk (Sweden)**. *Cancer Causes Control*, 1998. 9(3):259-267.

60. **Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease**. *Lancet*, 2001. 358(9291):1389-1399.

61. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, and 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.

62. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, and et al. **A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1**. *Science*, 1994. 266(5182):66-71.

63. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, and et al. **Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13**. *Science*, 1994. 265(5181):2088-2090.

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

65. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, and Micklem G. **Identification of the breast cancer susceptibility gene BRCA2**. *Nature*, 1995. 378(6559):789-792.

66. Thompson D and Easton D. **The genetic epidemiology of breast cancer genes**. *J Mammary Gland Biol Neoplasia*, 2004. 9(3):221-236.

67. Simard J, Tonin P, Durocher F, Morgan K, Rommens J, Gingras S, Samson C, Leblanc JF, Belanger C, Dion F, and et al. **Common origins of BRCA1 mutations in Canadian breast and ovarian cancer families**. *Nat Genet*, 1994. 8(4):392-398.

68. Martin AM and Weber BL. **Genetic and hormonal risk factors in breast cancer**. *J Natl Cancer Inst*, 2000. 92(14):1126-1135.

69. Berman DB, Costalas J, Schultz DC, Grana G, Daly M, and Godwin AK. **A common mutation in BRCA2 that predisposes to a variety of cancers is found in both Jewish Ashkenazi and non-Jewish individuals**. *Cancer Res*, 1996. 56(15):3409-3414.

70. Vehmanen P, Friedman LS, Eerola H, Sarantaus L, Pyrhonen S, Ponder BA, Muhonen T, and Nevanlinna H. **A low proportion of BRCA2 mutations in Finnish breast cancer families**. *Am J Hum Genet*, 1997. 60(5):1050-1058.

71. Thorlacius S, Olafsdottir G, Tryggvadottir L, Neuhausen S, Jonasson JG, Tavtigian SV, Tulinius H, Ogmundsdottir HM, and Eyfjord JE. **A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes**. *Nat Genet*, 1996. 13(1):117-119.

72. Bergman A, Flodin A, Engwall Y, Arkblad EL, Berg K, Einbeigi Z, Martinsson T, Wahlstrom J, Karlsson P, and Nordling M. **A high frequency of germline BRCA1/2 mutations in western Sweden detected with complementary screening techniques**. *Fam Cancer*, 2005. 4(2):89-96.

73. Szabo CI and King MC. **Population genetics of BRCA1 and BRCA2**. *Am J Hum Genet*, 1997. 60(5):1013-1020.

74. Arason A, Jonasdottir A, Barkardottir RB, Bergthorsson JT, Teare MD, Easton DF, and Egilsson V. **A population study of mutations and LOH at breast**

cancer gene loci in tumours from sister pairs: two recurrent mutations seem to account for all BRCA1/BRCA2 linked breast cancer in Iceland. *J Med Genet*, 1998. 35(6):446-449.

75. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, Struewing J, Arason A, Scherneck S, et al. **Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium**. *Am J Hum Genet*, 1998. 62(3):676-689.

76. Narod SA, Ford D, Devilee P, Barkardottir RB, Lynch HT, Smith SA, Ponder BA, Weber BL, Garber JE, Birch JM, and et al. **An evaluation of genetic heterogeneity in 145 breast-ovarian cancer families. Breast Cancer Linkage Consortium**. *Am J Hum Genet*, 1995. 56(1):254-264.

77. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, et al. **Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies**. *Am J Hum Genet*, 2003. 72(5):1117-1130.

78. Antoniou AC, Gayther SA, Stratton JF, Ponder BA, and Easton DF. **Risk models for familial ovarian and breast cancer**. *Genet Epidemiol*, 2000. 18(2):173-190.

79. Parkin DM, Pisani P, and Ferlay J. **Global cancer statistics**. *CA Cancer J Clin*, 1999. 49(1):33-64, 31.

80. **Cancer i siffror**, 2005: The National Board of Health and Welfare, Centre for Epidemiology, Official Statistics of Sweden.

81. Shiozawa T and Konishi I. **Early endometrial carcinoma: clinicopathology, hormonal aspects, molecular genetics, diagnosis, and treatment**. *Int J Clin Oncol*, 2006. 11(1):13-21.

82. Sherman ME, Sturgeon S, Brinton LA, Potischman N, Kurman RJ, Berman ML, Mortel R, Twiggs LB, Barrett RJ, and Wilbanks GD. **Risk factors and hormone levels in patients with serous and endometrioid uterine carcinomas**. *Mod Pathol*, 1997. 10(10):963-968.

83. Bokhman JV. **Two pathogenetic types of endometrial carcinoma**. *Gynecol Oncol*, 1983. 15(1):10-17.

84. Lax SF, Pizer ES, Ronnett BM, and Kurman RJ. **Clear cell carcinoma of the endometrium is characterized by a distinctive profile of p53, Ki-67, estrogen, and progesterone receptor expression**. *Hum Pathol*, 1998. 29(6):551-558.

85. Weiderpass E, Persson I, Adami HO, Magnusson C, Lindgren A, and Baron JA. **Body size in different periods of life, diabetes mellitus, hypertension, and risk of postmenopausal endometrial cancer (Sweden)**. *Cancer Causes Control*, 2000. 11(2):185-192.

86. Weiderpass E, Adami HO, Baron JA, Magnusson C, Bergstrom R, Lindgren A, Correia N, and Persson I. **Risk of endometrial cancer following estrogen replacement with and without progestins**. *J Natl Cancer Inst*, 1999. 91(13):1131-1137.

87. Hemminki K, Bermejo JL, and Granstrom C. **Endometrial cancer: population attributable risks from reproductive, familial and socioeconomic factors**. *Eur J Cancer*, 2005. 41(14):2155-2159.

88. Weiderpass E, Adami HO, Baron JA, Magnusson C, Lindgren A, and Persson I. **Use of oral contraceptives and endometrial cancer risk (Sweden)**. *Cancer Causes Control*, 1999. 10(4):277-284.

89. Weiderpass E and Baron JA. **Cigarette smoking, alcohol consumption, and endometrial cancer risk: a population-based study in Sweden**. *Cancer Causes Control*, 2001. 12(3):239-247.

90. Moradi T, Weiderpass E, Signorello LB, Persson I, Nyren O, and Adami HO. **Physical activity and postmenopausal endometrial cancer risk (Sweden)**. *Cancer Causes Control*, 2000. 11(9):829-837.

91. Kvale G, Heuch I, and Nilssen S. **Re: "Endometrial cancer and age at last delivery: evidence for an association"**. *Am J Epidemiol*, 1992. 135(4):453-455.

92. Aarnio M, Sankila R, Pukkala E, Salovaara R, Aaltonen LA, de la Chapelle A, Peltomaki P, Mecklin JP, and Jarvinen HJ. **Cancer risk in mutation carriers of DNA-mismatch-repair genes**. *Int J Cancer*, 1999. 81(2):214-218.

93. Vasen HF, Stormorken A, Menko FH, Nagengast FM, Kleibeuker JH, Griffioen G, Taal BG, Moller P, and Wijnen JT. **MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families**. *J Clin Oncol*, 2001. 19(20):4074-4080.

94. Hendriks YM, Wagner A, Morreau H, Menko F, Stormorken A, Quehenberger F, Sandkuijl L, Moller P, Genuardi M, Van Houwelingen H, Tops C, Van Puijenbroek M, Verkuijlen P, Kenter G, Van Mil A, et al. **Cancer risk in hereditary nonpolyposis colorectal cancer due to MSH6 mutations: impact on counseling and surveillance**. *Gastroenterology*, 2004. 127(1):17-25.

95. Carter J and Pather S. **An overview of uterine cancer and its management**. *Expert Rev Anticancer Ther*, 2006. 6(1):33-42.

96. Nakajin S, Shinoda M, Haniu M, Shively JE, and Hall PF. **C21 steroid side chain cleavage enzyme from porcine adrenal microsomes. Purification and characterization of the 17 alpha-hydroxylase/C17,20-lyase cytochrome P-450**. *J Biol Chem*, 1984. 259(6):3971-3976.

97. Picado-Leonard J and Miller WL. **Cloning and sequence of the human gene for P450c17 (steroid 17 alpha-hydroxylase/17,20 lyase): similarity with the gene for P450c21**. *DNA*, 1987. 6(5):439-448.

98. Carey AH, Waterworth D, Patel K, White D, Little J, Novelli P, Franks S, and Williamson R. **Polycystic ovaries and premature male pattern baldness are associated with one allele of the steroid metabolism gene CYP17**. *Hum Mol Genet*, 1994. 3(10):1873-1876.

99. Nedelcheva Kristensen V, Haraldsen EK, Anderson KB, Lonning PE, Erikstein B, Karesen R, Gabrielsen OS, and Borresen-Dale AL. **CYP17 and breast cancer risk: the polymorphism in the 5' flanking area of the gene does not influence binding to Sp-1**. *Cancer Res*, 1999. 59(12):2825-2828.

100. Chacko P, Rajan B, Mathew BS, Joseph T, and Pillai MR. **CYP17 and SULT1A1 gene polymorphisms in Indian breast cancer**. *Breast Cancer*, 2004. 11(4):380-388.

101. Miyoshi Y, Iwao K, Ikeda N, Egawa C, and Noguchi S. **Genetic polymorphism in CYP17 and breast cancer risk in Japanese women**. *Eur J Cancer*, 2000. 36(18):2375-2379.

102. Feigelson HS, Coetzee GA, Kolonel LN, Ross RK, and Henderson BE. **A polymorphism in the CYP17 gene increases the risk of breast cancer**. *Cancer Res*, 1997. 57(6):1063-1065.

103. Haiman CA, Hankinson SE, Spiegelman D, Colditz GA, Willett WC, Speizer FE, Kelsey KT, and Hunter DJ. **The relationship between a polymorphism in CYP17 with plasma hormone levels and breast cancer**. *Cancer Res*, 1999. 59(5):1015-1020.

104. Mitrunen K, Jourenkova N, Kataja V, Eskelinen M, Kosma VM, Benhamou S, Vainio H, Uusitupa M, and Hirvonen A. **Steroid metabolism gene CYP17 polymorphism and the development of breast cancer**. *Cancer Epidemiol Biomarkers Prev*, 2000. 9(12):1343-1348.

105. Ambrosone CB, Moysich KB, Furberg H, Freudenheim JL, Bowman ED, Ahmed S, Graham S, Vena JE, and Shields PG. **CYP17 genetic polymorphism, breast cancer, and breast cancer risk factors**. *Breast Cancer Res*, 2003. 5(2):R45-51.

106. Helzlsouer KJ, Huang HY, Strickland PT, Hoffman S, Alberg AJ, Comstock GW, and Bell DA. **Association between CYP17 polymorphisms and the development of breast cancer**. *Cancer Epidemiol Biomarkers Prev*, 1998. 7(10):945-949.

107. Huang CS, Chern HD, Chang KJ, Cheng CW, Hsu SM, and Shen CY. **Breast cancer risk associated with genotype polymorphism of the estrogen-metabolizing genes CYP17, CYP1A1, and COMT: a multigenic study on cancer susceptibility**. *Cancer Res*, 1999. 59(19):4870-4875.

108. Gudmundsdottir K, Thorlacius S, Jonasson JG, Sigfusson BF, Tryggvadottir L, and Eyfjord JE. **CYP17 promoter polymorphism and breast cancer risk in males and females in relation to BRCA2 status**. *Br J Cancer*, 2003. 88(6):933-936.

109. Weston A, Pan CF, Bleiweiss IJ, Ksieski HB, Roy N, Maloney N, and Wolff MS. **CYP17 genotype and breast cancer risk**. *Cancer Epidemiol Biomarkers Prev*, 1998. 7(10):941-944.

110. Hefler LA, Tempfer CB, Grimm C, Lebrecht A, Ulbrich E, Heinze G, Leodolter S, Schneeberger C, Mueller MW, Muendlein A, and Koelbl H. **Estrogen-metabolizing gene polymorphisms in the assessment of breast carcinoma risk and fibroadenoma risk in Caucasian women**. *Cancer*, 2004. 101(2):264-269.

111. Dunning AM, Dowsett M, Healey CS, Tee L, Luben RN, Folkerd E, Novik KL, Kelemen L, Ogata S, Pharoah PD, Easton DF, Day NE, and Ponder BA. **Polymorphisms associated with circulating sex hormone levels in postmenopausal women**. *J Natl Cancer Inst*, 2004. 96(12):936-945.

112. Wu AH, Seow A, Arakawa K, Van Den Berg D, Lee HP, and Yu MC. **HSD17B1 and CYP17 polymorphisms and breast cancer risk among Chinese women in Singapore**. *Int J Cancer*, 2003. 104(4):450-457.

113. Ahsan H, Whittemore AS, Chen Y, Senie RT, Hamilton SP, Wang Q, Gurvich I, and Santella RM. **Variants in estrogen-biosynthesis genes CYP17 and CYP19**

and breast cancer risk: a family-based genetic association study. *Breast Cancer Res*, 2005. 7(1):R71-81.

114. Verla-Tebit E, Wang-Gohrke S, and Chang-Claude J. **CYP17 5'-UTR MspA1 polymorphism and the risk of premenopausal breast cancer in a German population-based case-control study**. *Breast Cancer Res*, 2005. 7(4):R455-464.

115. Chang JH, Gertig DM, Chen X, Dite GS, Jenkins MA, Milne RL, Southey MC, McCredie MR, Giles GG, Chenevix-Trench G, Hopper JL, and Spurdle AB. **CYP17 genetic polymorphism, breast cancer, and breast cancer risk factors: Australian Breast Cancer Family Study**. *Breast Cancer Res*, 2005. 7(4):R513-521.

116. Ye Z and Parry JM. **The CYP17 MspA1 polymorphism and breast cancer risk: a meta-analysis**. *Mutagenesis*, 2002. 17(2):119-126.

117. Feigelson HS, McKean-Cowdin R, and Henderson BE. **Concerning the CYP17 MspA1 polymorphism and breast cancer risk: a meta-analysis**. *Mutagenesis*, 2002. 17(5):445-446; author reply 447-448.

118. Matsuoka S, Rotman G, Ogawa A, Shiloh Y, Tamai K, and Elledge SJ. **Ataxia telangiectasia-mutated phosphorylates Chk2 in vivo and in vitro**. *Proc Natl Acad Sci U S A*, 2000. 97(19):10389-10394.

119. Shieh SY, Ahn J, Tamai K, Taya Y, and Prives C. **The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites**. *Genes Dev*, 2000. 14(3):289-300.

120. Zeng Y, Forbes KC, Wu Z, Moreno S, Piwnica-Worms H, and Enoch T. **Replication checkpoint requires phosphorylation of the phosphatase Cdc25 by Cds1 or Chk1**. *Nature*, 1998. 395(6701):507-510.

121. Lee JS, Collins KM, Brown AL, Lee CH, and Chung JH. **hCds1-mediated phosphorylation of BRCA1 regulates the DNA damage response**. *Nature*, 2000. 404(6774):201-204.

122. Savitsky K, Bar-Shira A, Gilad S, Rotman G, Ziv Y, Vanagaite L, Tagle DA, Smith S, Uziel T, Sfez S, and et al. **A single ataxia telangiectasia gene with a product similar to PI-3 kinase**. *Science*, 1995. 268(5218):1749-1753.

123. Cortez D, Wang Y, Qin J, and Elledge SJ. **Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks**. *Science*, 1999. 286(5442):1162-1166.

124. Khosravi R, Maya R, Gottlieb T, Oren M, Shiloh Y, and Shkedy D. **Rapid ATM-dependent phosphorylation of MDM2 precedes p53 accumulation in response to DNA damage**. *Proc Natl Acad Sci U S A*, 1999. 96(26):14973-14977.

125. Olsen JH, Hahnemann JM, Borresen-Dale AL, Brondum-Nielsen K, Hammarstrom L, Kleinerman R, Kaariainen H, Lonnqvist T, Sankila R, Seersholm N, Tretli S, Yuen J, Boice JD, Jr., and Tucker M. **Cancer in patients with ataxia-telangiectasia and in their relatives in the nordic countries**. *J Natl Cancer Inst*, 2001. 93(2):121-127.

126. Swift M, Morrell D, Massey RB, and Chase CL. **Incidence of cancer in 161 families affected by ataxia-telangiectasia**. *N Engl J Med*, 1991. 325(26):1831-1836.

127. Thompson D, Duedal S, Kirner J, McGuffog L, Last J, Reiman A, Byrd P, Taylor M, and Easton DF. **Cancer risks and mortality in heterozygous ATM mutation carriers**. *J Natl Cancer Inst*, 2005. 97(11):813-822.

128. Inskip HM, Kinlen LJ, Taylor AM, Woods CG, and Arlett CF. **Risk of breast cancer and other cancers in heterozygotes for ataxia-telangiectasia**. *Br J Cancer*, 1999. 79(7-8):1304-1307.

129. Sommer SS, Jiang Z, Feng J, Buzin CH, Zheng J, Longmate J, Jung M, Moulds J, and Dritschilo A. **ATM missense mutations are frequent in patients with breast cancer**. *Cancer Genet Cytogenet*, 2003. 145(2):115-120.

130. Teraoka SN, Malone KE, Doody DR, Suter NM, Ostrander EA, Daling JR, and Concannon P. **Increased frequency of ATM mutations in breast carcinoma patients with early onset disease and positive family history**. *Cancer*, 2001. 92(3):479-487.

131. Chen J, Birkholtz GG, Lindblom P, Rubio C, and Lindblom A. **The role of ataxia-telangiectasia heterozygotes in familial breast cancer**. *Cancer Res*, 1998. 58(7):1376-1379.

132. FitzGerald MG, Bean JM, Hegde SR, Unsal H, MacDonald DJ, Harkin DP, Finkelstein DM, Isselbacher KJ, and Haber DA. **Heterozygous ATM mutations do not contribute to early onset of breast cancer**. *Nat Genet*, 1997. 15(3):307-310.

133. Concannon P. **ATM heterozygosity and cancer risk**. *Nat Genet*, 2002. 32(1):89-90.

134. Dork T, Bendix R, Bremer M, Rades D, Klopper K, Nicke M, Skawran B, Hector A, Yamini P, Steinmann D, Weise S, Stuhrmann M, and Karstens JH. **Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients**. *Cancer Res*, 2001. 61(20):7608-7615.

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

136. Chenevix-Trench G, Spurdle AB, Gatei M, Kelly H, Marsh A, Chen X, Donn K, Cummings M, Nyholt D, Jenkins MA, Scott C, Pupo GM, Dork T, Bendix R, Kirk J, et al. **Dominant negative ATM mutations in breast cancer families**. *J Natl Cancer Inst*, 2002. 94(3):205-215.

137. Stankovic T, Kidd AM, Sutcliffe A, McGuire GM, Robinson P, Weber P, Bedenham T, Bradwell AR, Easton DF, Lennox GG, Haites N, Byrd PJ, and Taylor AM. **ATM mutations and phenotypes in ataxia-telangiectasia families in the British Isles: expression of mutant ATM and the risk of leukemia, lymphoma, and breast cancer**. *Am J Hum Genet*, 1998. 62(2):334-345.

138. Larson GP, Zhang G, Ding S, Foldenauer K, Udar N, Gatti RA, Neuberg D, Lunetta KL, Ruckdeschel JC, Longmate J, Flanagan S, and Krontiris TG. **An allelic variant at the ATM locus is implicated in breast cancer susceptibility**. *Genet Test*, 1997. 1(3):165-170.

139. Szabo CI, Schutte M, Broeks A, Houwing-Duistermaat JJ, Thorstenson YR, Durocher F, Oldenburg RA, Wasielewski M, Odefrey F, Thompson D, Floore AN, Kraan J, Klijn JG, van den Ouweland AM, Wagner TM, et al. **Are ATM**

mutations 7271T-->G and IVS10-6T-->G really high-risk breast cancer-susceptibility alleles? *Cancer Res*, 2004. 64(3):840-843.

140. Buchholz TA, Weil MM, Ashorn CL, Strom EA, Sigurdson A, Bondy M, Chakraborty R, Cox JD, McNeese MD, and Story MD. **A Ser49Cys variant in the ataxia telangiectasia, mutated, gene that is more common in patients with breast carcinoma compared with population controls**. *Cancer*, 2004. 100(7):1345-1351.
141. Lee KM, Choi JY, Park SK, Chung HW, Ahn B, Yoo KY, Han W, Noh DY, Ahn SH, Kim H, Wei Q, and Kang D. **Genetic polymorphisms of ataxia telangiectasia mutated and breast cancer risk**. *Cancer Epidemiol Biomarkers Prev*, 2005. 14(4):821-825.
142. Bretsky P, Haiman CA, Gilad S, Yahalom J, Grossman A, Paglin S, Van Den Berg D, Kolonel LN, Skaliter R, and Henderson BE. **The relationship between twenty missense ATM variants and breast cancer risk: the Multiethnic Cohort**. *Cancer Epidemiol Biomarkers Prev*, 2003. 12(8):733-738.
143. Stredrick DL, Garcia-Closas M, Pineda MA, Bhatti P, Alexander BH, Doody MM, Lissowska J, Peplonska B, Brinton LA, Chanock SJ, Struewing JP, and Sigurdson AJ. **The ATM missense mutation p.Ser49Cys (c.146C>G) and the risk of breast cancer**. *Hum Mutat*, 2006. 27(6):538-544.
144. Angele S, Romestaing P, Moullan N, Vuillaume M, Chapot B, Friesen M, Jongmans W, Cox DG, Pisani P, Gerard JP, and Hall J. **ATM haplotypes and cellular response to DNA damage: association with breast cancer risk and clinical radiosensitivity**. *Cancer Res*, 2003. 63(24):8717-8725.
145. Koren M, Kimmel G, Ben-Asher E, Gal I, Papa MZ, Beckmann JS, Lancet D, Shamir R, and Friedman E. **ATM haplotypes and breast cancer risk in Jewish high-risk women**. *Br J Cancer*, 2006. 94(10):1537-1543.
146. Tamimi RM, Hankinson SE, Spiegelman D, Kraft P, Colditz GA, and Hunter DJ. **Common ataxia telangiectasia mutated haplotypes and risk of breast cancer: a nested case-control study**. *Breast Cancer Res*, 2004. 6(4):R416-422.
147. Scott SP, Bendix R, Chen P, Clark R, Dork T, and Lavin MF. **Missense mutations but not allelic variants alter the function of ATM by dominant interference in patients with breast cancer**. *Proc Natl Acad Sci U S A*, 2002. 99(2):925-930.
148. Bell DW, Varley JM, Szydlo TE, Kang DH, Wahrer DC, Shannon KE, Lubratovich M, Verselis SJ, Isselbacher KJ, Fraumeni JF, Birch JM, Li FP, Garber JE, and Haber DA. **Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome**. *Science*, 1999. 286(5449):2528-2531.
149. Li FP and Fraumeni JF, Jr. **Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome?** *Ann Intern Med*, 1969. 71(4):747-752.
150. Wu X, Webster SR, and Chen J. **Characterization of tumor-associated Chk2 mutations**. *J Biol Chem*, 2001. 276(4):2971-2974.
151. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, et al. **Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations**. *Nat Genet*, 2002. 31(1):55-59.

152. Vahteristo P, Bartkova J, Eerola H, Syrjakoski K, Ojala S, Kilpivaara O, Tamminen A, Kononen J, Aittomaki K, Heikkila P, Holli K, Blomqvist C, Bartek J, Kallioniemi OP, and Nevanlinna H. **A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer**. *Am J Hum Genet*, 2002. 71(2):432-438.

153. de Bock GH, Schutte M, Krol-Warmerdam EM, Seynaeve C, Blom J, Brekelmans CT, Meijers-Heijboer H, van Asperen CJ, Cornelisse CJ, Devilee P, Tollenaar RA, and Klijn JG. **Tumour characteristics and prognosis of breast cancer patients carrying the germline CHEK2*1100delC variant**. *J Med Genet*, 2004. 41(10):731-735.

154. Kilpivaara O, Bartkova J, Eerola H, Syrjakoski K, Vahteristo P, Lukas J, Blomqvist C, Holli K, Heikkila P, Sauter G, Kallioniemi OP, Bartek J, and 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(4):575-580.

155. Bogdanova N, Enbetaen-Dubrowinskaja N, Feshchenko S, Lazjuk GI, Rogov YI, Dammann O, Bremer M, Karstens JH, Sohn C, and Dork T. **Association of two mutations in the CHEK2 gene with breast cancer**. *Int J Cancer*, 2005. 116(2):263-266.

156. Kilpivaara O, Vahteristo P, Falck J, Syrjakoski K, Eerola H, Easton D, Bartkova J, Lukas J, Heikkila P, Aittomaki K, Holli K, Blomqvist C, Kallioniemi OP, Bartek J, and Nevanlinna H. **CHEK2 variant I157T may be associated with increased breast cancer risk**. *Int J Cancer*, 2004. 111(4):543-547.

157. Cybulski C, Gorski B, Huzarski T, Masojc B, Mierzejewski M, Debniak T, Teodorczyk U, Byrski T, Gronwald J, Matyjasik J, Zlowocka E, Lenner M, Grabowska E, Nej K, Castaneda J, et al. **CHEK2 is a multiorgan cancer susceptibility gene**. *Am J Hum Genet*, 2004. 75(6):1131-1135.

158. Gorski B, Cybulski C, Huzarski T, Byrski T, Gronwald J, Jakubowska A, Stawicka M, Gozdecka-Grodecka S, Szwiec M, Urbanski K, Mitus J, Marczyk E, Dziuba J, Wandzel P, Surdyka D, et al. **Breast cancer predisposing alleles in Poland**. *Breast Cancer Res Treat*, 2005. 92(1):19-24.

159. Dufault MR, Betz B, Wappenschmidt B, Hofmann W, Bandick K, Golla A, Pietschmann A, Nestle-Kramling C, Rhiem K, Huttner C, von Lindern C, Dall P, Kiechle M, Untch M, Jonat W, et al. **Limited relevance of the CHEK2 gene in hereditary breast cancer**. *Int J Cancer*, 2004. 110(3):320-325.

160. Schutte M, Seal S, Barfoot R, Meijers-Heijboer H, Wasielewski M, Evans DG, Eccles D, Meijers C, Lohman F, Klijn J, van den Ouweland A, Futreal PA, Nathanson KL, Weber BL, Easton DF, et al. **Variants in CHEK2 other than 1100delC do not make a major contribution to breast cancer susceptibility**. *Am J Hum Genet*, 2003. 72(4):1023-1028.

161. Friedrichsen DM, Malone KE, Doody DR, Daling JR, and Ostrander EA. **Frequency of CHEK2 mutations in a population based, case-control study of breast cancer in young women**. *Breast Cancer Res*, 2004. 6(6):R629-635.

162. Kuschel B, Auranen A, Gregory CS, Day NE, Easton DF, Ponder BA, Dunning AM, and Pharoah PD. **Common polymorphisms in checkpoint kinase 2 are not**

associated with breast cancer risk. *Cancer Epidemiol Biomarkers Prev*, 2003. 12(8):809-812.

163. Schechter AL, Stern DF, Vaidyanathan L, Decker SJ, Drebin JA, Greene MI, and Weinberg RA. **The neu oncogene: an erb-B-related gene encoding a 185,000-Mr tumour antigen**. *Nature*, 1984. 312(5994):513-516.

164. King CR, Kraus MH, and Aaronson SA. **Amplification of a novel v-erbB-related gene in a human mammary carcinoma**. *Science*, 1985. 229(4717):974-976.

165. Coussens L, Yang-Feng TL, Liao YC, Chen E, Gray A, McGrath J, Seeburg PH, Libermann TA, Schlessinger J, Francke U, and et al. **Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene**. *Science*, 1985. 230(4730):1132-1139.

166. Akiyama T, Sudo C, Ogawara H, Toyoshima K, and Yamamoto T. **The product of the human c-erbB-2 gene: a 185-kilodalton glycoprotein with tyrosine kinase activity**. *Science*, 1986. 232(4758):1644-1646.

167. Bargmann CI, Hung MC, and Weinberg RA. **The neu oncogene encodes an epidermal growth factor receptor-related protein**. *Nature*, 1986. 319(6050):226-230.

168. Threadgill DW, Dlugosz AA, Hansen LA, Tennenbaum T, Lichti U, Yee D, LaMantia C, Mourton T, Herrup K, Harris RC, and et al. **Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype**. *Science*, 1995. 269(5221):230-234.

169. Yarden Y and Sliwkowski MX. **Untangling the ErbB signalling network**. *Nat Rev Mol Cell Biol*, 2001. 2(2):127-137.

170. Lee KF, Simon H, Chen H, Bates B, Hung MC, and Hauser C. **Requirement for neuregulin receptor erbB2 in neural and cardiac development**. *Nature*, 1995. 378(6555):394-398.

171. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, and McGuire WL. **Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene**. *Science*, 1987. 235(4785):177-182.

172. Slamon DJ, Godolphin W, Jones LA, Holt JA, Wong SG, Keith DE, Levin WJ, Stuart SG, Udove J, Ullrich A, and et al. **Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer**. *Science*, 1989. 244(4905):707-712.

173. Revillion F, Bonneterre J, and Peyrat JP. **ERBB2 oncogene in human breast cancer and its clinical significance**. *Eur J Cancer*, 1998. 34(6):791-808.

174. Riben MW, Malfetano JH, Nazeer T, Muraca PJ, Ambros RA, and Ross JS. **Identification of HER-2/neu oncogene amplification by fluorescence in situ hybridization in stage I endometrial carcinoma**. *Mod Pathol*, 1997. 10(8):823-831.

175. Rolitsky CD, Theil KS, McGaughy VR, Copeland LJ, and Niemann TH. **HER-2/neu amplification and overexpression in endometrial carcinoma**. *Int J Gynecol Pathol*, 1999. 18(2):138-143.

176. Nazeer T, Ballouk F, Malfetano JH, Figge H, and Ambros RA. **Multivariate survival analysis of clinicopathologic features in surgical stage I**

endometrioid carcinoma including analysis of HER-2/neu expression. *Am J Obstet Gynecol*, 1995. 173(6):1829-1834.

177. Morrison C, Zanagnolo V, Ramirez N, Cohn DE, Kelbick N, Copeland L, Maxwell LG, and Fowler JM. **HER-2 is an independent prognostic factor in endometrial cancer: association with outcome in a large cohort of surgically staged patients**. *J Clin Oncol*, 2006. 24(15):2376-2385.
178. Benusiglio PR, Lesueur F, Luccarini C, Conroy DM, Shah M, Easton DF, Day NE, Dunning AM, Pharoah PD, and Ponder BA. **Common ERBB2 polymorphisms and risk of breast cancer in a white British population: a case-control study**. *Breast Cancer Res*, 2005. 7(2):R204-209.
179. Millikan RC, Hummer AJ, Wolff MS, Hishida A, and Begg CB. **HER2 codon 655 polymorphism and breast cancer: results from kin-cohort and case-control analyses**. *Breast Cancer Res Treat*, 2005. 89(3):309-312.
180. Cox DG, Hankinson SE, and Hunter DJ. **The erbB2/HER2/neu receptor polymorphism Ile655Val and breast cancer risk**. *Pharmacogenet Genomics*, 2005. 15(7):447-450.
181. Han W, Kang D, Lee JE, Park IA, Choi JY, Lee KM, Bae JY, Kim S, Shin ES, Shin HJ, Kim SW, and Noh DY. **A haplotype analysis of HER-2 gene polymorphisms: association with breast cancer risk, HER-2 protein expression in the tumor, and disease recurrence in Korea**. *Clin Cancer Res*, 2005. 11(13):4775-4778.
182. Yu Q, La Rose J, Zhang H, Takemura H, Kohn KW, and Pommier Y. **UCN-01 inhibits p53 up-regulation and abrogates gamma-radiation-induced G(2)-M checkpoint independently of p53 by targeting both of the checkpoint kinases, Chk2 and Chk1**. *Cancer Res*, 2002. 62(20):5743-5748.
183. Barlow C, Eckhaus MA, Schaffer AA, and Wynshaw-Boris A. **Atm haploinsufficiency results in increased sensitivity to sublethal doses of ionizing radiation in mice**. *Nat Genet*, 1999. 21(4):359-360.
184. Pietras RJ, Poen JC, Gallardo D, Wongvipat PN, Lee HJ, and Slamon DJ. **Monoclonal antibody to HER-2/neureceptor modulates repair of radiation-induced DNA damage and enhances radiosensitivity of human breast cancer cells overexpressing this oncogene**. *Cancer Res*, 1999. 59(6):1347-1355.
185. Liehr JG. **Genotoxicity of the steroidal oestrogens oestrone and oestradiol: possible mechanism of uterine and mammary cancer development**. *Hum Reprod Update*, 2001. 7(3):273-281.
186. Thibodeau PA, Kachadourian R, Lemay R, Bisson M, Day BJ, and Paquette B. **In vitro pro- and antioxidant properties of estrogens**. *J Steroid Biochem Mol Biol*, 2002. 81(3):227-236.
187. Ismail IH, Nystrom S, Nygren J, and Hammarsten O. **Activation of ataxia telangiectasia mutated by DNA strand break-inducing agents correlates closely with the number of DNA double strand breaks**. *J Biol Chem*, 2005. 280(6):4649-4655.
188. Brouillet JP, Dujardin MA, Chalbos D, Rey JM, Grenier J, Lamy PJ, Maudelonde T, and 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.

189. Isola J, DeVries S, Chu L, Ghazvini S, and Waldman F. **Analysis of changes in DNA sequence copy number by comparative genomic hybridization in archival paraffin-embedded tumor samples**. *Am J Pathol*, 1994. 145(6):1301-1308.

190. Pastinen T, Partanen J, and Syvanen AC. **Multiplex, fluorescent, solid-phase minisequencing for efficient screening of DNA sequence variation**. *Clin Chem*, 1996. 42(9):1391-1397.

191. Prince JA, Feuk L, Howell WM, Jobs M, Emahazion T, Blennow K, and Brookes AJ. **Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation**. *Genome Res*, 2001. 11(1):152-162.

192. Feigelson HS, Shames LS, Pike MC, Coetzee GA, Stanczyk FZ, and Henderson BE. **Cytochrome P450c17alpha gene (CYP17) polymorphism is associated with serum estrogen and progesterone concentrations**. *Cancer Res*, 1998. 58(4):585-587.

193. Sodha N, Williams R, Mangion J, Bullock SL, Yuille MR, and Eeles RA. **Screening hCHK2 for mutations**. *Science*, 2000. 289(5478):359.

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

195. Qin ZS, Niu T, and Liu JS. **Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms**. *Am J Hum Genet*, 2002. 71(5):1242-1247.

196. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, and 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.

197. Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, and 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.

198. Iles MM. **Obtaining unbiased estimates of tagging SNP performance**. *Ann Hum Genet*, 2006. 70(2):254-261.

199. Shoemaker J, Painter I, and Weir BS. **A Bayesian characterization of Hardy-Weinberg disequilibrium**. *Genetics*, 1998. 149(4):2079-2088.

200. Xu J, Turner A, Little J, Bleecker ER, and Meyers DA. **Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error?** *Hum Genet*, 2002. 111(6):573-574.

201. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, and Poland GA. **Score tests for association between traits and haplotypes when linkage phase is ambiguous**. *Am J Hum Genet*, 2002. 70(2):425-434.

202. Westfall P and Young S. **Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment**, 1993. New York: John Wiley & Sons, Inc.

203. Chapman JM, Cooper JD, Todd JA, and Clayton DG. **Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power**. *Hum Hered*, 2003. 56(1-3):18-31.

204. Gauderman WJ. **Sample size requirements for matched case-control studies of gene-environment interaction**. *Stat Med*, 2002. 21(1):35-50.

205. Manolio TA, Bailey-Wilson JE, and Collins FS. **Genes, environment and the value of prospective cohort studies**. *Nat Rev Genet*, 2006. 7(10):812-820.

206. Hernan MA, Hernandez-Diaz S, Werler MM, and Mitchell AA. **Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology**. *Am J Epidemiol*, 2002. 155(2):176-184.

207. Aschengrau A and Seage GR. **Essentials of Epidemiology in Public Health**, 2003. Sudbury, Massachusetts: Jones and Bartlett Publishers.

208. Garcia-Closas M, Rothman N, and Lubin J. **Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size**. *Cancer Epidemiol Biomarkers Prev*, 1999. 8(12):1043-1050.

209. Nystrom L, Larsson LG, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, Andersson I, Bjurstam N, Fagerberg G, Frisell J, and et al. **Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials. A comparison between official statistics and validation by an endpoint committee**. *Acta Oncol*, 1995. 34(2):145-152.

210. Pompanon F, Bonin A, Bellemain E, and Taberlet P. **Genotyping errors: causes, consequences and solutions**. *Nat Rev Genet*, 2005. 6(11):847-859.

211. Gordon D, Finch SJ, Nothnagel M, and Ott J. **Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms**. *Hum Hered*, 2002. 54(1):22-33.

212. Wacholder S, Rothman N, and Caporaso N. **Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer**. *Cancer Epidemiol Biomarkers Prev*, 2002. 11(6):513-520.

213. Pritchard JK and Rosenberg NA. **Use of unlinked genetic markers to detect population stratification in association studies**. *Am J Hum Genet*, 1999. 65(1):220-228.

214. Devlin B and Roeder K. **Genomic control for association studies**. *Biometrics*, 1999. 55(4):997-1004.

215. Pritchard JK, Stephens M, Rosenberg NA, and Donnelly P. **Association mapping in structured populations**. *Am J Hum Genet*, 2000. 67(1):170-181.

216. Cardon LR and Palmer LJ. **Population stratification and spurious allelic association**. *Lancet*, 2003. 361(9357):598-604.

217. **Immigration and emigration in the postwar period**, 2004: Statistics Sweden.

218. Bland JM and Altman DG. **Multiple significance tests: the Bonferroni method**. *Bmj*, 1995. 310(6973):170.

219. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, and Rothman N. **Assessing the probability that a positive report is false: an approach for molecular epidemiology studies**. *J Natl Cancer Inst*, 2004. 96(6):434-442.

220. Rothman K and Greenland S. **Causation and Causal Inference**. *Modern Epidemiology*, 1998. Lippincott-Raven Publishers: Philadelphia, PA.

221. Skrondal A. **Interaction as departure from additivity in case-control studies: a cautionary note**. *Am J Epidemiol*, 2003. 158(3):251-258.

222. Clayton D and McKeigue PM. **Epidemiological methods for studying genes and environmental factors in complex diseases**. *Lancet*, 2001. 358(9290):1356-1360.

223. Daneshmand S, Weitsman SR, Navab A, Jakimiuk AJ, and Magoffin DA. **Overexpression of theca-cell messenger RNA in polycystic ovary syndrome does not correlate with polymorphisms in the cholesterol side-chain cleavage and 17alpha-hydroxylase/C(17-20) lyase promoters**. *Fertil Steril*, 2002. 77(2):274-280.

224. Sharp L, Cardy AH, Cotton SC, and Little J. **CYP17 gene polymorphisms: prevalence and associations with hormone levels and related factors. a HuGE review**. *Am J Epidemiol*, 2004. 160(8):729-740.

225. Hong CC, Thompson HJ, Jiang C, Hammond GL, Tritchler D, Yaffe M, and Boyd NF. **Association between the T27C polymorphism in the cytochrome P450 c17alpha (CYP17) gene and risk factors for breast cancer**. *Breast Cancer Res Treat*, 2004. 88(3):217-230.

226. Tworoger SS, Chubak J, Aiello EJ, Ulrich CM, Atkinson C, Potter JD, Yasui Y, Stapleton PL, Lampe JW, Farin FM, Stanczyk FZ, and McTiernan A. **Association of CYP17, CYP19, CYP1B1, and COMT polymorphisms with serum and urinary sex hormone concentrations in postmenopausal women**. *Cancer Epidemiol Biomarkers Prev*, 2004. 13(1):94-101.

227. Garcia-Closas M, Herbstman J, Schiffman M, Glass A, and Dorgan JF. **Relationship between serum hormone concentrations, reproductive history, alcohol consumption and genetic polymorphisms in pre-menopausal women**. *Int J Cancer*, 2002. 102(2):172-178.

228. Gorai I, Tanaka K, Inada M, Morinaga H, Uchiyama Y, Kikuchi R, Chaki O, and Hirahara F. **Estrogen-metabolizing gene polymorphisms, but not estrogen receptor-alpha gene polymorphisms, are associated with the onset of menarche in healthy postmenopausal Japanese women**. *J Clin Endocrinol Metab*, 2003. 88(2):799-803.

229. Hamajima N, Iwata H, Obata Y, Matsuo K, Mizutani M, Iwase T, Miura S, Okuma K, Ohashi K, and Tajima K. **No association of the 5' promoter region polymorphism of CYP17 with breast cancer risk in Japan**. *Jpn J Cancer Res*, 2000. 91(9):880-885.

230. Lai J, Vesprini D, Chu W, Jernstrom H, and Narod SA. **CYP gene polymorphisms and early menarche**. *Mol Genet Metab*, 2001. 74(4):449-457.

231. Dunning AM, Healey CS, Pharoah PD, Foster NA, Lipscombe JM, Redman KL, Easton DF, Day NE, and Ponder BA. **No association between a polymorphism in the steroid metabolism gene CYP17 and risk of breast cancer**. *Br J Cancer*, 1998. 77(11):2045-2047.

232. Goodman MT, McDuffie K, Guo C, Terada K, and Donlon TA. **CYP17 genotype and ovarian cancer: a null case-control study**. *Cancer Epidemiol Biomarkers Prev*, 2001. 10(5):563-564.

233. Feigelson HS, McKean-Cowdin R, Pike MC, Coetzee GA, Kolonel LN, Nomura AM, Le Marchand L, and Henderson BE. **Cytochrome P450c17alpha gene (CYP17) polymorphism predicts use of hormone replacement therapy**. *Cancer Res*, 1999. 59(16):3908-3910.

234. Haiman CA, Hankinson SE, Colditz GA, Hunter DJ, and De Vivo I. **A polymorphism in CYP17 and endometrial cancer risk**. *Cancer Res*, 2001. 61(10):3955-3960.

235. McKean-Cowdin R, Feigelson HS, Pike MC, Coetzee GA, Kolonel LN, and Henderson BE. **Risk of endometrial cancer and estrogen replacement therapy history by CYP17 genotype**. *Cancer Res*, 2001. 61(3):848-849.

236. de Jong MM, Nolte IM, Te Meerman GJ, van der Graaf WT, Oosterom E, Bruinenberg M, Steege G, Oosterwijk JC, van der Hout AH, Boezen HM, Schaapveld M, Kleibeuker JH, and de Vries EG. **No increased susceptibility to breast cancer from combined CHEK2 1100delC genotype and the HLA class III region risk factors**. *Eur J Cancer*, 2005. 41(12):1819-1823.

237. Mateus Pereira LH, Sigurdson AJ, Doody MM, Pineda MA, Alexander BH, Greene MH, and Struewing JP. **CHEK2:1100delC and female breast cancer in the United States**. *Int J Cancer*, 2004. 112(3):541-543.

238. Offit K, Pierce H, Kirchhoff T, Kolachana P, Rapaport B, Gregersen P, Johnson S, Yossepowitch O, Huang H, Satagopan J, Robson M, Scheuer L, Nafa K, and Ellis N. **Frequency of CHEK2*1100delC in New York breast cancer cases and controls**. *BMC Med Genet*, 2003. 4(1):1.

239. Osorio A, Rodriguez-Lopez R, Diez O, de la Hoya M, Ignacio Martinez J, Vega A, Esteban-Cardenosa E, Alonso C, Caldes T, and Benitez J. **The breast cancer low-penetrance allele 1100delC in the CHEK2 gene is not present in Spanish familial breast cancer population**. *Int J Cancer*, 2004. 108(1):54-56.

240. Caligo MA, Agata S, Aceto G, Crucianelli R, Manoukian S, Peissel B, Scaini MC, Sensi E, Veschi S, Cama A, Radice P, Viel A, D'Andrea E, and Montagna M. **The CHEK2 c.1100delC mutation plays an irrelevant role in breast cancer predisposition in Italy**. *Hum Mutat*, 2004. 24(1):100-101.

241. Huzarski T, Cybulski C, Domagala W, Gronwald J, Byrski T, Szwiec M, Woyke S, Narod SA, and Lubinski J. **Pathology of breast cancer in women with constitutional CHEK2 mutations**. *Breast Cancer Res Treat*, 2005. 90(2):187-189.

242. Goode EL, Dunning AM, Kuschel B, Healey CS, Day NE, Ponder BA, Easton DF, and Pharoah PP. **Effect of germ-line genetic variation on breast cancer survival in a population-based study**. *Cancer Res*, 2002. 62(11):3052-3057.

243. Cavalieri EL, Rogan EG, and Chakravarti D. **Initiation of cancer and other diseases by catechol ortho-quinones: a unifying mechanism**. *Cell Mol Life Sci*, 2002. 59(4):665-681.

244. Tanko LB and Christiansen C. **An update on the antiestrogenic effect of smoking: a literature review with implications for researchers and practitioners**. *Menopause*, 2004. 11(1):104-109.

245. Westhoff C, Gentile G, Lee J, Zacur H, and Helbig D. **Predictors of ovarian steroid secretion in reproductive-age women**. *Am J Epidemiol*, 1996. 144(4):381-388.

246. Sterzik K, Strehler E, De Santo M, Trumpp N, Abt M, Rosenbusch B, and Schneider A. **Influence of smoking on fertility in women attending an in vitro fertilization program**. *Fertil Steril*, 1996. 65(4):810-814.

247. Van Voorhis BJ, Dawson JD, Stovall DW, Sparks AE, and Syrop CH. **The effects of smoking on ovarian function and fertility during assisted reproduction cycles**. *Obstet Gynecol*, 1996. 88(5):785-791.

248. Persson I. **Estrogens in the causation of breast, endometrial and ovarian cancers - evidence and hypotheses from epidemiological findings**. *J Steroid Biochem Mol Biol*, 2000. 74(5):357-364.

249. Hill AB. **The Environment and Disease: Association or Causation?** *Proc R Soc Med*, 1965. 58:295-300.

250. Zhang J and Powell SN. **The role of the BRCA1 tumor suppressor in DNA double-strand break repair**. *Mol Cancer Res*, 2005. 3(10):531-539.

251. Wang HC, Chou WC, Shieh SY, and Shen CY. **Ataxia telangiectasia mutated and checkpoint kinase 2 regulate BRCA1 to promote the fidelity of DNA end-joining**. *Cancer Res*, 2006. 66(3):1391-1400.