Published in final edited form as:

Breast Cancer Res Treat. 2007 April ; 102(2): 237–247. doi:10.1007/s10549-006-9324-7.

A Comprehensive Examination of CYP19 Variation and Risk of Breast Cancer Using Two Haplotype Tagging Approaches

Janet E. Olson¹, James N. Ingle², Cynthia X. Ma³, Linda L. Pellegymounter⁴, Daniel J. Schaid⁵, V. Shane Pankratz⁵, Robert A. Vierkant⁵, Zachary S. Fredericksen⁵, Yanhong Wu⁶, Fergus Couch⁷, Celine M. Vachon¹, Thomas A. Sellers⁸, and Richard M. Weinshilboum⁴

¹Department of Health Sciences Research, Division of Epidemiology, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55901

²Department of Oncology, Division of Medical Oncology, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55901

³Division of Oncology, Siteman Comprehensive Cancer Center, Barnes-Jewish Hospital and Washington University School of Medicine, Box 8100, 660 South Euclid Avenue; St. Louis, MO 3110–1093

⁴Department of Molecular Pharmacology and Experimental Therapeutics, Division of Clinical Pharmacology, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55901

⁵Department of Health Sciences Research, Division of Biostatistics, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN, 55901

⁶Genotyping Shared Resource, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN 55901

⁷Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, 200 First St. SW, Rochester, MN, 55901,

⁸Division of Cancer Prevention and Control, H. Lee Moffitt Cancer Center & Research Institute, 12902 Magnolia Drive, Tampa, FL, 33612

Abstract

Background—Numerous studies point to a positive relationship between elevated levels of estrogens and increased risk of breast. Androgens are converted to estrogens by the aromatase enzyme, which is encoded by the CYP19 gene. We recently published resequencing data on 88 polymorphisms identified in that gene. The hypothesis tested in this study was that polymorphisms, or haplotypes, in CYP19 are related to risk of breast cancer.

Methods—Incident cases of breast cancer were identified through the Division of Medical Oncology at the Mayo Clinic in Rochester, MN. Controls were patients visiting Mayo for an annual medical examination. Controls were frequency matched to cases based on age and region of residence. Tag-polymorphisms were selected using 2 methods: 1) 12 variants using the tag-selection method of Carlson et al [1]; and 2) 12 variants using the haplotype method of Stram [2]. Six SNPs were selected by both methods. Genotyping was conducted using SNPstream, TaqMan and RFLP analyses. Logistic regression was used to calculate odds ratios (OR) and 95% confidence intervals

(CI). Analyses were conducted among all cases and controls, or stratified by estrogen receptor alpha (ER) status and/or menopausal status.

Results—A total of 750 cases (60% postmenopausal) and 732 controls (75% postmenopausal) were included. No association with breast cancer risk was detected for individual variants, selected tagSNPs or hap-tag SNPs despite 80% power to detect odds ratios as low as 1.49 for MAF of 0.10. Similarly, stratified analyses based on ER status or menopausal status failed to detect any association with breast cancer risk.

Conclusion—These analyses suggest that variants of *CYP19* are not associated with risk of breast cancer.

Keywords

Aromatase; Breast; Breast Cancer; CYP19; CYP19A1; Epidemiology; Etiology; Molecular Biology

Introduction

There is strong evidence for a relationship between estrogens and breast cancer [3,4]. Many studies have found that elevated levels of estrogens are associated with increase risk of breast cancer [3]. Obesity, which is associated with higher levels of estrogen in postmenopausal women, has been consistently linked with increased risk of postmenopausal breast cancer [5]. Early menarche, late first full-term pregnancy, and late menopause, which are all associated with higher lifetime exposures to estrogens, have also been consistently reported to increase risk for breast cancer [4,6,7]. The ovaries are the primary source of estrogens in the premenopausal woman whereas in postmenopausal women estrogens arise from the conversion of androgens to estrogens in peripheral tissues including skin, fat, muscle, etc by the aromatase enzyme, which is encoded by the *CYP19* gene. The importance of the aromatase enzyme is illustrated by the importance of aromatase inhibitors in clinical practice for treatment of postmenopausal women with estrogen receptor positive advanced breast cancer [8] and early stage disease [9].

Although there is an indisputable relationship between estrogens and breast cancer, the pathophysiological mechanism(s) is unclear. Estrogens have been thought to exert their carcinogenic actions through interaction with the estrogen receptor (ER). A complementary mechanism has been proposed in which estrogens can be metabolized to genotoxic metabolites that interact with DNA to form depurinating adducts and cause apurinic sites in DNA that are subject to error-prone repair resulting in mutations [10,11].

Epidemiological studies of common *CYP19* polymorphisms have generated inconsistent results in regard to the possible association with risk for breast cancer. The first reports focused on a polymorphic tetranucleotide repeat (TTTA)_n in intron 4. Initially, a 2.4-fold elevated risk of breast cancer in individuals carrying the longest repeat (TTTA)₁₂ compared to those carrying the other repeats ($p < 0.05$) was observed [12]. Subsequent studies have reported increased risk associated with the (TTTA)₁₀ allele [13–15], the (TTTA)₈ allele [15], and the (TTTA)₇ allele [16]. However, two nested case-control studies from large cohorts reported no association of the (TTTA)_n variation with breast cancer risk [17,18]. Complicating interpretation of these findings is the lack of consistency in the number of repeats comprising the referent (low risk) allele. This fact, and its location within an intron, suggest that any association with risk of breast cancer would likely be due to linkage disequilibrium to another causal variant in the vicinity.

Other variations in *CYP19* have also been examined with regard to possible risk of breast cancer. A C to T substitution in the 3' untranslated region (UTR) of exon 10 of the *CYP19*

gene (rs10046) was detected twice as commonly in 481 cases as in 236 controls (OR 2.00; 95% CI, 1.28 – 3.11)[19]. However, these findings were not confirmed in a nested study within the Nurses' Health Study[20]. Hirose et al examined a non-synonymous coding SNP (W39R) in a case-control study of Japanese women and found no association between the SNP and risk of breast cancer (OR 1.21; 95% CI 0.69–2.14). However, homozygous and heterozygous carriers of the variant allele did show a significantly increased risk of breast cancer among premenopausal women with a late age at first full-term pregnancy (OR 7.31, 1.88–28.5) or premenopausal women with a high body mass index (OR 2.77, 1.12–6.87). Other polymorphisms studied with relation to risk of breast cancer include a TTC deletion in intron 4 and a polymorphism at codon 264 [17]. Neither of these were associated with increased risk of breast cancer [17].

Each of these studies examined only a portion of the total genetic variation in the *CYP19* gene. We recently published a report of a systematic resequencing of all coding exons, upstream untranslated exons and core promoter regions, all exon-intron splice junctions and a portion of the 3'-untranslated region of *CYP19* [21]. A total of 88 polymorphisms were identified in the 240 samples examined, 37 of these in Caucasian samples. Here we report the results of a comprehensive analysis of the contribution of these 37 polymorphisms to the risk of breast cancer in 750 cases and 732 controls. We systematically selected representative SNPs from the gene using two methods [1,2] to determine whether a specific method provided greater insight into the gene and disease association. The hypothesis tested in this study was that polymorphisms, or haplotypes, in the gene encoding the aromatase enzyme are associated with risk of breast cancer. This hypothesis was studied in both pre-and post-menopausal women as the tissue levels of estrogens in the breast might be expected to be influenced by aromatase irrespective of the menopausal status.

Methods

Population

The Mayo Clinic Breast Cancer Case-Control study is an on-going, unselected, clinic-based series of breast cancer cases and healthy controls. Breast cancer patients were women diagnosed within the previous six months with no prior history of cancer (except non-melanoma skin cancer) who were seen in the Division of Medical Oncology between February 1, 2001 and June 2005. Controls were selected from women without a prior history of cancer (except non-melanoma skin cancer) visiting Mayo Clinic for a general medical exam in the Department of Internal Medicine. Controls were frequency matched to cases on region of residence and five-year age group. All subjects used in this study were restricted to those with residence in the states of Iowa, Minnesota, Wisconsin, North or South Dakota, or Illinois. Cases and controls were mostly Caucasian (94% and 92%, respectively). All eligible women were asked to provide informed consent, risk factor information via a written questionnaire, and a sample of blood as a source of DNA. The questionnaire collected information on height, and weight, physical activity (moderate vs. vigorous; frequency), alcohol intake, multivitamin intake, ages of menses and menopause, use of exogenous hormones (years taken, year last taken, formulation), reproductive history, history of oophorectomy and/or hysterectomy, breast biopsy history, detailed family history, education and race/ethnicity. Education was categorized into groups of equal to, less than or greater than a high school education. Postmenopausal was defined as having no menstrual period for 12 months or having had their uterus and/or ovaries removed. Hormone replacement therapy was asked as a three-level variable: currently taking, taken in the past or never used. The composition of hormone replacement therapy was identified as estrogen alone, estrogen and Progesterone combination or unknown. Family history of cancer was reported for all first and second degree relatives,

and included information on type of cancer and age of onset. Smoking status including environmental and personal smoking history.

Laboratory Methods

Selection of variants

SNPs that captured the genetic variability in the gene were selected from all variants identified through the resequencing of the gene. As reported earlier [21], DNA samples from 60 Caucasian-American, 60 African-American, 60 Han Chinese-American, and 60 Mexican-American subjects were obtained from the Coriell Cell Repository (Camden, NJ) and resequenced.

This included all upstream untranslated exons plus their core promoter regions, all exon-intron splice junctions and a portion of the 3'-untranslated region of *CYP19* [21]. Only the data from the 60 Caucasian-American samples were used for SNP selection for the present analyses. Two subsets of the variants that explained most of the genetic variability in the gene were selected. Parameters for the haplotype-tagging approach of Stram [2] were: 0.02 minimum minor allele frequency (MAF), a threshold of 1% for minimum haplotype frequency, and a minimum r^2 of 0.90. This resulted in 39 haplotypes. Parameters for the LD Select method of Carlson [1] were: a minimum MAF of 0.05, and 80% correlation within bins (See table 2). Twenty six loci had a MAF of 0.05 or more. Hereafter, the method of Stram will be referred to as the HT-Tag method; Carlson's method will be referred to as the Linkage Disequilibrium (LD-Tag) method. The variants identified through either of the tagging approaches were genotyped in the breast cancer cases and controls. The genotypes in this series of subjects provided the venue in which the desired evaluation of a potential breast cancer association with the *CYP19* gene took place.

High-throughput multiplex PCR and SNP analysis

The majority of the genotyping was performed with GenomeLab SNPstream genotyping platform (Beckman Coulter) and its accompanying SNPstream software suite. The 12 pairs of primers for the multiplex PCR and single-base extension primers were designed for the specific SNPs and flanking sequences in a homogenous reaction with Web-based software provided at <http://www.autoprimer.com/> (Beckman Coulter Inc. Fullerton, CA). SNPstream genotyping was done as previously described [22,23]. In brief, the SNPs of interest were amplified in a 12-pair multiplex PCR under universal conditions (5mM MgCl₂, 75μM dNTPs, 0.1 unit AmpliTaq Gold (Applied Biosystems) in a final volume of 5 μl; PCR 94°C 1 minute; 34 cycles 94°C 30 seconds, 55°C 30 seconds, 72°C 1 minute). Single base extension primers were extended with single TAMRA- or Bodipy-fluorescein-labeled nucleotide terminator reactions (96°C 3 minutes, then 45 cycles of 94°C 20 seconds, 40°C 11 seconds) and then spatially resolved by hybridization to the complementary oligonucleotides arrayed on the 384-well (SNPware Tag array) microplates. The 12 individual SNPs were identified by their position and fluorescent color in each well according to the position of the tagged oligonucleotides. Genotype data were generated on the basis of the relative fluorescent intensities for each SNP. Graphical review and operator adjustment of the genotype clusters were performed to refine fluorescent cutoff values.

Other genotyping

Genotyping of the repeat length polymorphisms, Intron 4 (77) (TTTA)_n and the Intron 4 (27) TCT insertion/deletion were carried out using an ABI 3730 DNA sequencer and analyzed using GeneMapper software (Applied Biosystems, Foster City, CA). The Exon I.6 SNP (−77) G>A (rs4775936), was genotyped using the TaqMan methodology (Perkin-Elmer ABI Prism 7700 Sequence Detection System; Applied Biosystems, Foster City, CA). PCR were carried out in reaction volumes of 25 μL containing 50 ng DNA, 1× TaqMan Universal PCR Master Mix

(Applied Biosystems), 280 nmol/L of each primer, 6-carboxyfluorescein-labeled minor groove binder probe, and VIC-labeled minor groove binder probe. PCR conditions included an initial incubation at 50°C for 2 minutes, 95°C for 10 minutes followed by 35 cycles of 15 seconds at 92°C and 1 minute at 60°C. Genotype analysis was performed using the ABI Prism 7700 SDS software (Applied Biosystems).

The quality of the genotyping was assessed by estimation of Hardy Weinberg Equilibrium among controls. Blinded duplicates of 13 samples were included in each genotyping run. Most duplicates were concordant or discordant for only one. Exon 1.2 (–224) had two discordant duplicates and should be interpreted with caution.

Statistical Analysis

Analysis of Risk of Breast Cancer

The genotypes obtained from the study subjects were used to estimate allele frequencies. In analyses with the Intron (TTTA)_n VNTR, the alleles with $n=7, 9, 11$, or 13 were defined as the referent category because these repeat lengths had never been previously associated with risk of breast cancer in previous scientific reports (reviewed in the Introduction).

Estimates of pair-wise linkage disequilibrium, both D' and r^2 , were also obtained using the data from control subjects. Following these initial evaluations of the variants, associations between each individual variant and breast cancer case-control status were performed. Single-variant analyses were performed using logistic regression, where case-control status was the response, and genotypes were modeled as having a log-additive relationship with breast cancer case status. Thus, analyses estimate the change in the odds of breast cancer associated with each additional copy of the variant allele. All models adjusted for the design variables of age and behavioral variables could confound the association between the SNPs of interest and breast cancer, we examined the independent effects of the following variables on breast cancer risk using a backward elimination selection approach: education, age at menarche, number of live births, age at first live birth, fertility problems, oral contraceptive use, menopausal status, HRT use, family history of breast or ovarian cancer in first or second degree relatives, smoking status, and body mass index. All variables associated with case-control status in the final stepwise model ($p<0.05$) were considered potential confounders and included in all subsequent SNP analyses.

Overall differences in risk of breast cancer across haplotypes were assessed using a global score test of Schaid et al [24]. After these global tests were performed, we examined individual haplotype effects. Such comparisons were performed in the spirit of Fisher's protected least significant difference test: individual associations were not considered to be statistically significant in the absence of global significance. Individual haplotype associations compared the haplotype of interest against all other haplotypes combined.

Separate models were fit for LD-Tag SNPs and HT-Tag SNPs. Primary analyses were conducted among all cases and controls, but secondary analyses stratified by estrogen receptor alpha (ER) status, breast cancer stage, and/or menopausal status. Analyses were performed in the SAS (SAS Institute, Cary NC) and S-Plus (Insightful, Seattle WA) statistical packages.

Results

The Mayo Breast Cancer Case-Control study included 750 cases overall; 591 invasive, and 159 ductal carcinoma *in situ*. Of these, 462 (74.6%) were estrogen receptor positive, 110 (17.8%) were estrogen receptor negative, and the remainder were missing ER status. A similar

proportion (70%) were positive for the progesterone receptor (N=429), whereas 141 (22.8%) were PR negative, and the remainder were untested for PR status.

Table 1 shows demographic characteristics of cases and controls included in the analyses. Cases were slightly younger than controls (56.4 vs. 58.0; $p=0.01$). Approximately 45% of the cases lived in counties in Southeastern Minnesota at the time of diagnosis. The remainder of cases resided in other counties of Minnesota (22%), Iowa (18%), Wisconsin (11%), or the Dakotas/Illinois (3%) at the time of diagnosis. Similar proportions of controls were from the same regions. More controls (N=331, 46.8%) than cases (N=285, 38.5%) reported being “highly” physically active in the past year, $p<0.01$. Cases had a higher body mass index (BMI), a younger age at menarche, older age at menopause, and a stronger family history of cancer than controls. The majority of cases (N=454, 63.2%) were postmenopausal at the time of accrual into the study. Seventy-three percent of the controls were past menopause at the time of accrual. The participation rate among cases was 75% and 72% among controls.

Twelve LD bins or groups of SNPs correlated by LD were estimated by the LD-Tag method of Carlson [1], with a minimum minor allele frequency (MAF) of 0.05 and 80% correlation within bins (Table 2). Similarly, twelve variants were selected based upon the HT-Tag method of Stram [2] with a threshold MAF of 0.02 and a minimum haplotype frequency of 0.01 and a minimum R^2 for predicting haplotypes of 90%. Six SNPs were common across the two methods. In addition, two non-synonymous cSNPs were also genotyped: W39R and R264C. T201M was selected as a tagging SNP in the LD-tag method.

Table 3 displays the odds ratios (OR's) for the association between each *CYP19* variant and breast cancer after adjustment for age and region of residence only or adjustment for all significant risk factors in a multivariate model. None of the variants were associated with elevated risk of breast cancer. The non-synonymous cSNPs (W39R, R264C and T201M) were also not associated with risk of breast cancer. Based upon the hypothesis that risk of breast cancer often varies by menopausal status, we examined the association between the SNPs and breast cancer when stratifying cases and controls by menopausal status or by whether tumors were estrogen receptor alpha positive. No changes in estimates of disease risk were observed (data not shown). In other analyses, cases were limited to only invasive breast cancer and the analyses repeated. Again, the results were unchanged (data not shown).

The variants selected by the Stram [2] method were designed to “tag” or represent variants with which there was 90% or greater correlation. Similarly, tagSNPs chosen by the LD-Tag method of Carlson [1] were representative of SNPs with 80% correlation by LD in LD bins (Table 2). Thus we expected that the HT-Tag selected variants would be correlated to each other by less than 80%, and the variants selected by Carlson's LD-Tag method would be correlated to one another by less than 80%. However, the figures show that more than the expected correlation was evident between some of the variants. Therefore, we conducted haplotype-based analyses, stratified by the selection method (LD-tag vs HT-tag).

Table 4 and Table 5 contain the analyses of estimated *CYP19* haplotypes in cases and their controls. Seventeen haplotypes were detected among subjects when examining the variants selected with the HT-tag method of Stram [2] with a haplotype frequency of 0.02 or greater. Twelve haplotypes meeting the same criteria existed in subjects when analyzing variants selected with the LD-tag method of Carlson [1]. One haplotype in each group (Table 4, Haplotype #17; Table 5, Haplotype #1) was significantly associated with increased risk of breast cancer when compared to all other haplotypes combined, (Stram HT-tag, $p=0.049$; Carlson LD-tag $p=0.014$). However, the whole gene tests of association between haplotypes and breast cancer status was not statistically significant in either analyses ($p=0.36$, $p=0.10$, respectively). When cases were restricted to only those with estrogen receptor positive (ER+)

tumors, the LD-tag approach of Carlson yielded a marginally significant global gene test, $p=0.051$. Two haplotypes (identical to haplotypes #1 and #12 in table 5) were individually associated with risk of breast cancer. Haplotype #1 was associated with decreased risk of breast cancer ($p=0.033$), and haplotype #12 was associated with increased risk of breast cancer ($p=0.044$). These two haplotypes differ by only one SNP in the 5' flanking region of exon 1.1 (-588), which was itself not associated with risk in the single SNP analyses. Therefore the importance of such an association is unclear. Similar analyses were conducted in postmenopausal women and premenopausal women (data not shown). None of the global tests for these sets of analyses reached statistical significance.

Four nonsynonymous coding SNPs were identified in our earlier resequencing [21]. One of these, M364T, is extremely rare and has been reported only in Asian populations. Therefore, it was not genotyped in our study. The W39R was genotyped, but only one individual carried a variant allele. The frequencies of the T201M and R264C non-synonymous cSNPs were 0.03 and 0.03 in cases and 0.03 and 0.04 in controls, respectively. In haplotype analyses of these two cSNPs, no individuals carried a copy of both variants. Ninety-three percent of the haplotypes were Thr²⁰¹ Arg²⁶⁴, 0.031 were Met²⁰¹ Arg²⁶⁴, and 0.038 were Thr²⁰¹ Cys²⁶⁴. None of these haplotypes were associated with risk of breast cancer overall or when stratified by menopausal status, or limited to only ER positive tumors. cSNPs (W39R, R264C and T201M)

Power Considerations

The total sample size consisted of a total of 750 breast cancer cases and 732 controls. This sample size provided power to detect significant associations of reasonably small size, especially for allele frequencies above 0.10. Using the power formulae developed for the Armitage test for trend [25], we computed the minimum odds ratio that would be detected with 80% power with a type I error rate of 5% with a Bonferroni correction for multiple testing. We had at least 80% power to detect odds ratios of 1.5 or higher for allele frequencies of 10% or greater. For the maximum observed allele frequency (44%), we had power to detect odds ratios of 1.3 or greater and for the minimum observed allele frequency (0.02), we had power to detect odds ratios of 2.14 or larger (See Table 6).

Discussion

The aromatase enzyme, coded by the *CYP19* gene, is vitally important for normal hormone function. Previous reports of the relationship of *CYP19* variants and risk of breast cancer have not been comprehensive and have not yielded a clear picture of its potential role in development of the disease. We have taken a haplotype-tagging approach to variant selection, such that all common variation within the gene has been studied with relationship to risk of breast cancer. We did this in a clinic-based case-control study of breast cancer in 756 cases and 726 matched controls.

Tagging methods are an efficient way to select and genotype a maximally informative set of common variants within a candidate gene [1,2]. Complete assay of all the variation within a gene is not usually feasible, nor is it statistically efficient, because many variants are correlated with each other. In addition, some variants may be so rare that very few subjects within a study would be expected to carry it, making statistical conclusions impossible. Two commonly used methods of selecting representative SNPs from a gene are the method of Carlson [1] and that of Stram [2]. One of the major differences between the two methods is that the Carlson method does not require variants within a haplotype bin to lie contiguously on the strand of DNA. Rather, location is ignored, and the correlation between pairs of SNPs takes precedence. Both methods disregard the more rare variants. Our analyses selected two sets of variants for study:

one set of twelve selected using the haplotype-tag method of Stram and another set of twelve selected using the LD-tag method of Carlson and colleagues. Six variants were in common across the two methods. We used MAF cutoffs of 0.05 for the Carlson method and 0.02 for the Stram method. Variants less common than this were not examined in our analysis. No association was detected for any variants, selected by either method, with regard to risk of breast cancer in our population. In addition, no association with risk of breast cancer was detected for any haplotype in either method.

The first manuscripts published on CYP19 variants and risk of breast cancer examined the variable number of tandem repeats (VNTR) located in intron 4. Several reports [12–16] indicated a possible association with risk of breast cancer; however, because of the intronic location, it was considered likely to be due to linkage disequilibrium with a functional site within the gene or gene region. Within our resequencing data, this variation was in high LD with nine other variants, none of which are responsible for a codon change. Five of the nine variants were also selected by the HT-Tag method, were genotyped and were examined directly for association with risk of breast cancer. The other three were located in Exon 3 (A240G), IVS5-16T>G, and IVS7-79A>G. Therefore, if the previously published results were truly due to linkage disequilibrium with a functional site located elsewhere, there is no evidence that the site was in CYP19. This VNTR was not associated with increased risk of breast cancer in our population. No association was detected when cases were limited to only those whose tumors were estrogen receptor positive, when stratified on menopausal status, or when limited to only invasive cases of cancer. Similarly, we did not detect any association with risk of breast cancer for non-synonymous coding SNPs, or for variants in the untranslated region (UTR) of exon 10. We examined both the C to T substitution (rs10046) previously reported to be linked to breast cancer by Kristensen [19] as well as a G to T substitution located 142 basepairs away (rs4646).

Another of our goals within this study was to evaluate whether the method of SNP tag selection influenced the results. Therefore, we systematically selected representative SNPs from the gene using two methods [1,2] to determine whether a specific method provided greater insight into the gene and disease association. There were no differences in the scientific conclusions reached within our study. All of the variants selected by either method pointed to a lack of association between these variants and risk of breast cancer. We did, however, notice that the method of Carlson [1] was more variable in the number of SNPs as we modified the SNP selection parameters. For example, when we set the parameters to a minimum MAF of 0.02 (previously 0.05), and 90% (previously 80%) correlation within bins, the number of variants required to represent the majority of variation changed from 12 to 21.

One of the lessons from our analysis was that the residual correlation between haplotype tagging variants justifies the analysis of risk for disease by common haplotypes. We had earlier wondered whether analysis of risk by haplotype would be necessary. However, each tag selection method selected a group of tag SNPs that were strongly correlated, making haplotype analysis relevant. Two of our variants selected by the LD-tag method ((-628) and (596)) were 84% correlated. This is higher than expected as it exceeds the 80% correlation parameter used in the selection process. One possible reason for this is that the population on which our tag-selection process was conducted was not a subset of our own population. Rather, we used 60 Caucasian samples from the Coriell Cell Repository (Camden, NJ). This emphasizes the importance of checking residual correlations between SNPs within studies of this type. Study designs that use data from small groups of subjects available from public databases to select tag SNPs are at even greater risk of selecting tag SNPs that do not perfectly represent the variation within their own study population.

A major strength of this study is the complete gene resequencing data from which we selected our haplotype tagging SNPs. This allowed us to investigate all of the common variation within this gene. Previous studies interrogated only a portion of the genetic variation within this gene. As mentioned earlier rare variants were not examined. Another strength of this study is the reasonably large population of breast cancer cases and controls within which we examined our hypotheses. With the sample size available, we had power to detect significant associations of 1.5 or greater, especially for allele frequencies above 10%. For the alleles on the lower bound of frequency (0.05), only odds ratios of 1.69 or larger would have been detected.

A limitation of this study is the clinic-based nature of the cases and controls. Compared to the Iowa SEER data, our cases are somewhat younger than all breast cancer cases in the Iowa Registry. This is somewhat expected, as women who are very old are unlikely to travel to a tertiary center for care of breast cancer. Ductal carcinoma *in situ* is slightly more common in Mayo cases than in Iowa SEER, but there were no strong differences between the Mayo cases and the Iowa SEER cases for tumor stage and ER status. From this comparison, we believe that the breast cancer cases in our sample are reasonably representative of all the breast cancer cases in the general population of this area. Women included in this study were over 90% Caucasian, therefore these data are not generalizable to other population groups.

Another limitation of this study was the limited power to examine gene-environment interactions. Although we have thoroughly tested for evidence of the main effects of *CYP19* genetic polymorphisms with regard to risk of breast cancer in this population, there may still exist interactions with environmental factors that have not been detected. Therefore we encourage others with larger sample sizes to examine the joint effects of this gene and environmental factors.

In summary, we conducted a case-control analysis of breast cancer using a comprehensive selection of tagging variants to represent the majority of the common variants in the entire *CYP19* gene. We also examined risk among cases stratified by both menopausal status and limited to only those cases that were estrogen receptor positive. We found no evidence that variation in *CYP19* is associated with risk of breast cancer.

Acknowledgments

This work was supported by grants from the National Cancer Institute P01 CA 82267, P50 CA116201 and the National Institute of General Medical Sciences R01 GM28157, R01 GM35720, and U01 GM 61388 (The Pharmacogenetics Research Network)

References

1. Carlson CS, Eberle MA, et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004;74:106–120. [PubMed: 14681826]
2. Stram DO. Tag SNP selection for association studies. *Genet Epidemiol* 2004;27:365–374. [PubMed: 15372618]
3. Yager JD, Davidson NE. Estrogen carcinogenesis in breast cancer. *N Engl J Med* 2006;354:270–282. [PubMed: 16421368]
4. Clemons M, Goss P. Estrogen and the risk of breast cancer. *N Engl J Med* 2001;344:276–285. [PubMed: 11172156]
5. Key TJ, Appleby PN, et al. Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women. *J Natl Cancer Inst* 2003;95:1218–1226. [PubMed: 12928347]
6. Rosner B, Colditz GA. Nurses' health study: log-incidence mathematical model of breast cancer incidence. *J Natl Cancer Inst* 1996;88:359–364. [PubMed: 8609645]

7. Paffenbarger RS Jr, Kampert JB, et al. Characteristics that predict risk of breast cancer before and after the menopause. *Am J Epidemiol* 1980;112:258–268. [PubMed: 7416152]
8. Ingle JN, Suman VJ. Aromatase inhibitors for therapy of advanced breast cancer. *J Steroid Biochem Mol Biol* 2005;95:113–119. [PubMed: 15939585]
9. Ingle JN. Adjuvant endocrine therapy for postmenopausal women with early breast cancer. *Clin Cancer Res* 2006;12:1031s–1036s. [PubMed: 16467121]
10. Cavalieri E, Frenkel K, et al. Estrogens as endogenous genotoxic agents--DNA adducts and mutations. *J Natl Cancer Inst Monogr* 2000;75–93. [PubMed: 10963621]
11. Santen RJ.; Yue, W., et al. Estradiol-induced carcinogenesis via formation of genotoxic metabolites; *Advances in Endocrine Therapy of Breast Cancer: Proceedings of the 2003 Gleneagles Conference*; New York, NY: Marcel Dekker; 2004. p. 13-177.
12. Kristensen VN, Andersen TI, et al. A rare CYP19 (aromatase) variant may increase the risk of breast cancer. *Pharmacogenetics* 1998;8:43–48. [PubMed: 9511180]
13. Haiman CA, Hankinson SE, et al. A tetranucleotide repeat polymorphism in CYP19 and breast cancer risk. *Int J Cancer* 2000;87:204–210. [PubMed: 10861475]
14. Siegelmann-Danieli N, Buetow KH. Constitutional genetic variation at the human aromatase gene (Cyp19) and breast cancer risk. *Br J Cancer* 1999;79:456–463. [PubMed: 10027313]
15. Baxter SW, Choong DY, et al. Polymorphic variation in CYP19 and the risk of breast cancer. *Carcinogenesis* 2001;22:347–349. [PubMed: 11181459]
16. Miyoshi Y, Ando A, et al. Association of genetic polymorphisms in CYP19 and CYP1A1 with the oestrogen receptor-positive breast cancer risk. *Eur J Cancer* 2003;39:2531–2537. [PubMed: 14602139]
17. Probst-Hensch NM, Ingles SA, et al. Aromatase and breast cancer susceptibility. *Endocr Relat Cancer* 1999;6:165–173. [PubMed: 10731105]
18. Thyagarajan B, Brott M, et al. CYP1B1 and CYP19 gene polymorphisms and breast cancer incidence: no association in the ARIC study. *Cancer Lett* 2004;207:183–189. [PubMed: 15072827]
19. Kristensen VN, Harada N, et al. Genetic variants of CYP19 (aromatase) and breast cancer risk. *Oncogene* 2000;19:1329–1333. [PubMed: 10713674]
20. Haiman CA, Hankinson SE, et al. No association between a single nucleotide polymorphism in CYP19 and breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 2002;11:215–216. [PubMed: 11867511]
21. Ma CX, Adjei AA, et al. Human aromatase: gene resequencing and functional genomics. *Cancer Res* 2005;65:11071–11082. [PubMed: 16322257]
22. Bell PA, Chaturvedi S, et al. SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Biotechniques* 2002;70–72. 74, 76–77. [PubMed: 12083401]
23. Denomme GA, Van Oene M. High-throughput multiplex single-nucleotide polymorphism analysis for red cell and platelet antigen genotypes. *Transfusion* 2005;45:660–666. [PubMed: 15847652]
24. Schaid DJ, Rowland CM, et al. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;70:425–434. [PubMed: 11791212]
25. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11:375–386.

Table 1

Comparisons of demographic and breast cancer risk factors in the Mayo Clinic Breast Cancer Case and Control Study, 2001 to 2005

Variable	Level	Breast Cancer Cases (N=750) ^a	Controls (N=732) ^a	P-Value ^b
Age (years)	Mean (N, S.D.)	56.4 (750, 12.01)	58 (732, 12.25)	0.01
Geographic Area	Olmsted	144 (19.2%)	127 (17.3%)	0.90
	SE Minnesota	200 (26.7%)	202 (27.6%)	
	Minnesota other	166 (22.1%)	163 (22.3%)	
	Wisconsin	80 (10.7%)	71 (9.7%)	
	Iowa	136 (18.1%)	147 (20.1%)	
	Other Dakotas	24 (3.2%)	22 (3.0%)	
Physical activity	Low	202 (27.3%)	183 (25.8%)	<0.01
	Medium	253 (34.2%)	194 (27.4%)	
	High	285 (38.5%)	331 (46.8%)	
Body Mass Index (kg/m ²)	Mean (N, S.D.)	27.6 (734, 6.14)	27 (716, 5.81)	0.07
Age at Menarche (years)	Missing	46 (6%)	57 (6%)	0.05
	< 12	126 (17.9%)	113 (16.7%)	
	12	208 (29.5%)	160 (23.7%)	
	13	205 (29.1%)	219 (32.4%)	
	14+	165 (23.4%)	183 (27.1%)	
Menopausal Status	Premenopausal	264 (36.8%)	193 (27%)	<0.01
	Postmenopausal	454 (63.2%)	521 (73%)	
Age at Menopause	Mean (N, S.D.)	48.1 (454, 6.96)	46.3 (482, 6.36)	<0.01
Oral Contraceptive use	Never	259 (35.6%)	238 (34.2%)	0.49
	1–48 Months	190 (26.1%)	170 (24.4%)	
	48+ Months	279 (38.3%)	288 (41.4%)	
HRT	Never	384 (54.5%)	337 (51.1%)	0.25
	1–60 Months	140 (19.9%)	128 (19.4%)	
	60+ Months	180 (25.6%)	195 (29.5%)	
Number of live births	nulliparous	95 (12.8%)	113 (15.8%)	<0.01
	1–2	320 (43.2%)	250 (35%)	
	3+	325 (43.9%)	351 (49.2%)	
Age at first live birth	Nulliparous	95 (12.9%)	113 (15.9%)	0.10
	≤20	171 (23.2%)	139 (19.6%)	
	20+	471 (63.9%)	457 (64.5%)	
Family History of Breast or Ovarian Cancer	No	374 (50.4%)	422 (58.9%)	<0.01
	Yes	368 (49.6%)	295 (41.1%)	
Smoking Status	Never	452 (61.1%)	442 (61.6%)	0.82
	Ever	288 (38.9%)	275 (38.4%)	

Variable	Level	Breast Cancer Cases (N=750) ^a	Controls (N=732) ^a	P-Value ^b
Alcohol Use	Never	95 (12.9%)	96 (13.5%)	0.05
	Monthly	334 (45.5%)	276 (38.7%)	
	Weekly	231 (31.5%)	267 (37.4%)	
	Daily	74 (10.1%)	74 (10.4%)	
Education	No diploma	43 (5.8%)	26 (3.6%)	0.06
	HS diploma	187 (25.4%)	166 (23.2%)	
	post HS education	506 (68.8%)	524 (73.2%)	

^aValues presented as number (percent) unless otherwise indicated

^bP-values based on chi-square tests for categorical variables and t-tests for continuous variables.

Variants selected for genotyping using the methods outlined by Sram (Haplotype-tagging) and Carlson(Linkage Disequilibrium)

Table 2

	RS ID #	Location	NT	Seq. Change	Amino Acid Change	Frequency in Coriell Caucasians	MAF Freq in Mayo Breast Cancer Controls
Haplotype-tagging							
1 HT1	rs6493497	5'FR Exon I.1	-144	C-->T		0.16	0.11
2 HT2	<i>q</i> _{rs4774585}	5'FR Exon 2a	-468	C-->T		0.18	0.17
HT3	none	5'FR Exon 2a	-429	T-->C		0.04	0.04
4 HT4	<i>q</i> _{rs936308}	5'FR Exon I.5	-628	C-->G		0.87	0.84
5 HT5	<i>q</i> _{rs2008691}	5'FR Exon I.2	-596	T-->C		0.13	0.16
6 HT6	rs1062033	Exon I.2	-224	G-->C		0.45	0.46
7 HT7	<i>q</i> _{rs10459592}	5'FR Exon I.6	-196	A-->C		0.61	0.57
8 HT8	rs4775936	Exon I.6	-77	G-->A		0.49	0.43
9 HT9	<i>q</i> _{rs3759811}	Intron 2	-59	A-->G		0.54	0.5
10 HT10	VNTR	Intron 4	77	(TTTA) <i>n</i>			
11 HT11	rs10046	3'UTR	1531	C-->T		0.56	0.49
12 HT12	<i>q</i> _{rs4646}	3'UTR	1673	G-->T		0.29	0.27
Linkage Disequilibrium							
1 LD1	rs7176005	5'FR Exon I.1	-588	G-->A		0.14	0.11
2 LD2	rs7181886	Intron I.7	54	G-->C		0.06	0.05
3 LD3	none	5'FR Exon I.f	-725	G-->A		0.09	0.11
4 LD4	<i>q</i> _{rs3759811}	Intron 2	-59	A-->G		0.54	0.5
5 LD5	<i>q</i> _{rs4774585}	5'FR Exon 2a	-468	C-->T		0.18	0.02
6 LD6	<i>q</i> _{rs936308}	5'FR Exon I.5	-628	C-->G		0.87	0.84
7 LD7	<i>q</i> _{rs2008691}	5'FR Exon I.2	-596	T-->C		0.13	0.16
8 LD8	<i>q</i> _{rs10459592}	5'FR Exon I.6	-196	A-->C		0.61	0.57
9 LD9	rs11575899	Intron 4	27	TCT I/D		0.33	0.33
10 LD10	None	Exon 5	602	C-->T	Thr (201) Met	0.05	0.03
11 LD11	rs2289105	Intron 7	26	C-->T		0.10	0.08

RS ID #	Location	NT	Seq. Change	Amino Acid Change	Frequency in Coriell Caucasians	MAF Freq in Mayo Breast Cancer Controls
12 LD12 <i>a</i> _{rs4646}	3'UTR	1673	G->T		0.29	0.27

a variant selected by both methods

Table 3

Association of variants of CYP19 selected by haplotype-tagging methods of Carlson and/or Stram,

CYP19 Variant	Minor Allele Frequency		Age & Region Adjusted		Multivariate Adjusted ^a	
	Cases	Controls	Odds Ratio ^b	95% Confidence Interval	Odds Ratio ^b	95% Confidence Interval
5'FR Exon 1.1 (-588)	0.11	0.11	1.04	(0.83 – 1.31)	1.06	(0.83 – 1.35)
5'FR Exon 1.1 (-144)	0.12	0.11	1.10	(0.87 – 1.39)	1.11	(0.87 – 1.42)
5'FR Exon 2.a (-468)	0.19	0.18	1.14	(0.94 – 1.39)	1.20	(0.97 – 1.47)
5'FR Exon 2.a (-429)	0.03	0.04	0.77	(0.52 – 1.15)	0.77	(0.51 – 1.17)
5'FR Exon 1.5 (-628)	0.15	0.16	0.96	(0.79 – 1.18)	0.94	(0.76 – 1.17)
Intron 1.7 (54)	0.06	0.05	1.06	(0.77 – 1.44)	1.09	(0.79 – 1.52)
5'FR Exon 1.f (-725)	0.10	0.11	0.89	(0.70 – 1.13)	0.86	(0.67 – 1.10)
5'FR Exon 1.2 (-596)	0.15	0.16	0.92	(0.75 – 1.13)	0.92	(0.74 – 1.14)
Exon 1.2 (224)	0.47	0.46	1.06	(0.92 – 1.22)	1.03	(0.89 – 1.20)
5'FR Exon 1.6 (-196)	0.42	0.43	0.97	(0.84 – 1.13)	1.00	(0.85 – 1.16)
Exon 1.6 (-77)	0.43	0.43	1.02	(0.88 – 1.19)	1.01	(0.87 – 1.17)
Intron 2 (59)	0.48	0.50	0.95	(0.83 – 1.10)	0.97	(0.84 – 1.13)
Intron 4 (27) I/D	0.34	0.34	1.03	(0.88 – 1.20)	1.02	(0.88 – 1.19)
Intron 4 (VNTR) ⁿ c	0.85	0.84	1.00		1.00	
(VNTR)8	0.11	0.11	0.97	(0.76 – 1.23)	0.90	(0.70 – 1.15)
(VNTR)10	0.01	0.01	0.93	(0.47 – 1.83)	1.08	(0.52 – 2.24)
(VNTR)12	0.03	0.03	1.09	(0.70 – 1.68)	1.23	(0.78 – 1.94)
Exon 5 (602) (Thr ²⁰¹ Met)	0.03	0.03	1.06	(0.69 – 1.63)	1.02	(0.65 – 1.61)
Intron 7 (26)	0.09	0.08	1.02	(0.78 – 1.33)	1.04	(0.79 – 1.37)
3'UTR Exon 10 (1531)	0.47	0.49	0.95	(0.82 – 1.09)	0.98	(0.84 – 1.14)
3'UTR Exon 10 (1673)	0.25	0.27	0.90	(0.76 – 1.06)	0.91	(0.77 – 1.09)

^aAdjusted for age, region of residence at entry into study, age at menarche, menopause status, education, activity, family history of breast and/or ovarian cancer, and alcohol intake

^bThe odds ratio represents the estimated increase in the odds of breast cancer for each extra copy of the minor allele carried by the individual

^cReference category for (VNTR)_n analysis included alleles in which n=7, 9, 11 or 13

Within the Mayo Clinic Breast Cancer Case Control Study

Table 4
Associations of CYP19 haplotypes with risk of breast cancer based on SNPs selected using the HT-tag approach of Stram^a

Hap #	5'FR Exon 1.1 (-144)	5'FR Exon 2a (-468)	5'FR Exon 2a (-429)	5'FR Exon 1.5 (-628)	5'FR Exon 1.2 (-596)	5'FR Exon 1.2 (-224)	5'FR Exon 1.6 (-77)	Intron 2 (-59)	Intron 4 (TTTA)n	3'UTR Exon 10 (1531)	3'UTR Exon 10 (1673)	Hap. Freq. ^b	Score Test ^c	p-Value ^d
1	C	C	T	G	T	G	C	T	LR ^e	C	T	0.020	-1.728	0.084
2	C	C	T	C	C	G	A	T	LR ^e	C	T	0.085	-1.533	0.125
3	C	C	T	G	T	C	C	T	344	T	G	0.012	-1.457	0.145
4	C	C	T	G	T	C	C	T	328	T	G	0.081	-1.116	0.264
5	C	C	T	G	T	G	A	T	LR ^e	C	T	0.066	-0.800	0.424
6	T	C	T	G	T	G	A	C	LR ^e	T	G	0.026	-0.628	0.530
7	C	C	T	C	C	G	A	T	LR ^e	C	G	0.036	-0.370	0.711
8	C	C	C	G	T	C	C	T	LR ^e	T	G	0.026	-0.155	0.877
9	C	C	T	G	T	C	C	T	LR ^e	T	G	0.161	-0.074	0.941
10	C	C	T	G	T	C	C	C	LR ^e	T	G	0.025	0.246	0.806
11	C	C	T	G	T	C	C	C	328	T	G	0.011	0.506	0.613
12	C	T	T	G	T	G	C	T	LR ^e	C	T	0.045	0.607	0.544
13	T	C	T	G	T	G	A	T	LR ^e	C	G	0.030	0.761	0.447
14	C	C	T	G	T	G	A	T	LR ^e	C	G	0.115	0.847	0.397
15	C	T	T	G	T	C	C	T	LR ^e	T	G	0.078	0.908	0.364
16	T	C	T	G	T	G	C	T	LR ^e	C	T	0.019	1.034	0.301
17	C	T	T	G	T	C	C	T	344	T	G	0.014	1.971	0.049

^aResults adjust for the effects of age, region of residence at entry into study, age at menarche, menopause status, education, and activity, family history of breast and/or ovarian cancer, and alcohol intake. Selection parameters: 0.02 minor allele frequency, 1% minimum haplotype frequency, and a minimum r^2 of 0.90. Global test of association: $p=0.36$.

^bEstimated haplotype frequency

^cScore statistic comparing the haplotype of interest to all other haplotypes combined. Negative values imply decreased risk of breast cancer, positive values imply increased risk.

^dP-value comparing the haplotype of interest to all other haplotypes combined.

^eThese alleles are those never reported in the scientific literature to be associated with risk of breast cancer (TTTA)n, n=7, 9, 11, or 13

Table 5
Associations of CYP19 haplotypes with risk of breast cancer based on SNPs selected using the LD-tag approach of Carlson^a

Hap #	5'FR Exon 1.1 (-588)	5'FR Exon 2a (-468)	5'FR Exon 1.5 (-628)	Intron 1.7 54	5'FR Exon 1f (-725)	5'FR Exon 1.2 (-596)	5'FR Exon 1.6 (-196)	Intron 2 (-59)	Intron 4 27	Exon 5 602	Intron 7 26	3'UTR 1673	Hap. Freq. ^b	Score Test ^c	P Value ^d
1	G	C	G	G	G	T	C	T	INS	C	T	T	0.018	-2.468	0.014
2	G	C	C	G	A	C	A	T	DEL	C	C	T	0.081	-1.215	0.224
3	G	C	C	C	G	C	A	T	INS	C	C	G	0.037	-1.202	0.229
4	G	C	G	G	G	T	A	T	DEL	C	C	T	0.063	-1.001	0.317
5	G	C	G	G	G	T	C	C	INS	C	C	G	0.335	-0.810	0.418
6	A	C	G	G	G	T	A	T	DEL	C	C	G	0.030	-0.432	0.666
7	A	C	G	G	G	T	A	C	INS	T	C	G	0.027	-0.088	0.930
8	A	C	G	G	G	T	C	C	INS	C	C	G	0.022	0.193	0.847
9	G	T	G	G	G	T	C	T	INS	C	T	T	0.044	0.329	0.743
10	G	C	G	G	G	T	A	T	DEL	C	C	G	0.120	1.110	0.267
11	G	T	G	G	G	T	C	C	INS	C	C	G	0.114	1.452	0.147
12	A	C	G	G	G	T	C	T	INS	C	T	T	0.019	1.500	0.133

^aResults adjust for the effects of age, region of residence at entry into study, age at menarche, menopause status, education, activity, family history of breast and/or ovarian cancer, and alcohol intake. Selection parameters: 0.05 minor allele frequency, 80% correlation within bins. Global test of association: p=0.10.

^bEstimated haplotype frequency

^cScore statistic comparing the haplotype of interest to all other haplotypes combined. Negative values imply decreased risk of breast cancer, positive values imply increased risk.

^dP-value comparing the haplotype of interest to all other haplotypes combined.

Table 6

Minimum detectable odds ratios for an additive association between SNP and breast cancer, with type I error set at 5% and corrected for 12 independent tests.

Allele frequency (%)	80% Power	90% Power
2	2.14	2.30
5	1.69	1.78
10	1.49	1.55
20	1.36	1.41
30	1.32	1.36
40	1.30	1.34
44	1.30	1.34