*Human Molecular Genetics, 2003, Vol. 12, No. 20* **2679–2692**
*DOI: 10.1093/hmg/ddg294*

# A comprehensive haplotype analysis of *CYP19* and breast cancer risk: the Multiethnic Cohort

**Christopher A. Haiman[1,*], Daniel O. Stram[1], Malcolm C. Pike[1], Laurence N. Kolonel[2], Noel P. Burtt[3], David Altshuler[3,4,5,6], Joel Hirschhorn[3,6,7] and Brian E. Henderson[1]**

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA, [2]Cancer Etiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, HI 96813, USA, [3]Whitehead/MIT Center for Genome Research, Cambridge, MA 02139, USA, [4]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA, [5]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA, [6]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA and [7]Division of Endocrinology, Children's Hospital and Department of Pediatrics, Boston, MA 02115, USA

Received June 18, 2003; Revised and Accepted August 16, 2003

**The *CYP19* gene encodes for aromatase (P450arom), a key steroidogenic enzyme that catalyzes the final step of estrogen biosynthesis. Apart from rare mutations in *CYP19* which result in severe phenotypes associated with estrogen insufficiency, little is known about whether common variation in *CYP19* is associated with risk of hormone-related diseases. In this study, we employed a haplotype-based approach to search for common disease-associated variants in this candidate breast cancer susceptibility gene among African-American, Hawaiian, Japanese, Latina and White women in the Multiethnic Cohort Study (MEC). We utilized 74 densely spaced single-nucleotide polymorphisms (SNPs) (one every ~2.6 kb) spanning 189.4 kb of the *CYP19* locus to characterize linkage disequilibrium (LD) and haplotype patterns among 69–70 individuals from each ethnic population. We detected four regions of strong LD (blocks 1–4) that were quite closely conserved across populations. Within each block there was a limited diversity of common haplotypes (5 to 10 with a frequency ≥5%) and most haplotypes were observed to be shared across populations. Twenty-five haplotype-tagging SNPs (htSNPs) were selected to predict the common haplotypes with high probability (average $R_h^2 = 0.92$) and genotyped in a breast cancer case–control study in the MEC (cases, $n = 1355$; controls, $n = 2580$). We first performed global tests for differences in risk according to the common haplotypes and observed significant haplotype-effects in block 2 [$P = 0.01$; haplotypes 2b (OR = 1.23; 95% CI, 1.07–1.40), 2d (OR = 1.28; 95% CI, 1.01–1.62)]. We also found a common long-range haplotype comprised of block-specific haplotypes 2b and 3c to be associated with increased risk of breast cancer (haplotype 2b–3c: OR = 1.31; 95% CI, 1.11–1.54). Our findings suggest the hypothesis that women with the long-range *CYP19* haplotype 2b–3c may be carriers of a predisposing breast cancer susceptibility allele.**

## INTRODUCTION

Estrogens stimulate breast cell division and have an established role in breast carcinogenesis (1). Among postmenopausal woman, greater endogenous estrogen levels have been consistently associated with increased breast cancer risk (2,3). Prior to menopause, estrogens are primarily produced in the ovaries, while among postmenopausal women most circulating estrogens are synthesized from adrenal androgens in adipose tissue. $C_{19}$ androgens, androstenedione and testosterone, are converted to $C_{18}$ estrogens, estrone and estradiol, respectively, by the cytochrome P450 enzyme, aromatase. In humans, aromatase is expressed in the gonads as well as various other extragonadal sites, including adipose, placenta, skin, brain and bone. Aromatase is encoded by the *CYP19* gene which is located at 15q21.1 and spans ~123 kb. The gene comprises nine coding exons (II–X) covering ~30 kb, with multiple untranslated first exons localized within ~90 kb 5′ of

*To whom correspondence should be addressed at: Department of Preventive Medicine, University of Southern California, USC/Norris Comprehensive Cancer Center, 1441 Eastlake Ave, Rm 4441, Los Angeles, CA 90089-9175, USA. Tel: +1 3238650429; Fax: +1 3238650127; Email: haiman@usc.edu

the coding region that are regulated by tissue-specific promoters (4–6).

Studies suggest that aromatase has a direct effect on *in situ* estrogen synthesis in the breast and implicate the transcriptional regulation of *CYP19* in the development and progression of breast cancer (7–10). Among postmenopausal women, estradiol levels in malignant breast tissue have been observed to be higher than in non-malignant breast tissue and in the circulation (11). Elevated levels of aromatase expression have also been observed in breast tumors and adjacent tissue, relative to normal breast tissue (8). The heightened aromatase expression is accompanied by a change in *CYP19* promoter utilization, from the adipose-specific glucocorticoid-stimulated promoter I.4 to proximal promoter II which drives aromatase expression in the ovary and promoter I.3, which is a minor promoter used in adipose tissue (12,13).

The importance of the aromatase enzyme in the pathogenesis of breast cancer has also been clearly demonstrated in the clinical setting, as steroidal and nonsteroidal inhibitors of the enzyme have been used as second-line therapy following tamoxifen treatment for postmenopausal women with advanced breast cancer (14). Current studies also suggest that aromatase inhibitors (letrozole and anastrozole) may be equally or more effective than modulators of the estrogen receptor in slowing tumor progression, and support their use as first-line treatment for women with hormone receptor-positive breast cancer (15–17).

Aside from mutations in key breast cancer susceptibility genes BRCA1 and BRCA2, which are highly penetrant but explain only a relatively small percentage of breast cancer in the general population (<5%), the genetic risk factors contributing to sporadic breast cancer are as yet not known. Based on the evidence implicating aromatase in the underlying pathogenesis of breast cancer, we selected *CYP19* as a candidate gene to evaluate in relationship with breast cancer risk. Previous studies evaluating genetic variation in *CYP19* have examined relatively few polymorphic sites. The most well-studied polymorphism is the tetranucleotide $(TTTA)_n$ repeat in intron 4, but for the most part, associations between specific repeat alleles and breast cancer risk have been inconsistent (18–22). No comprehensive study of the role of this gene in breast cancer has been performed.

Haplotype-based association studies have been proposed as a powerful comprehensive approach to identify causal genetic variation underlying complex diseases (23,24). Recently, studies have shown that the human genome is comprised of genomic segments (blocks) that display little evidence of historical recombination and low haplotype diversity (23–25). Due to the high degree of linkage disequilibrium (LD) observed between single-nucleotide polymorphisms (SNPs) within these blocks, ancestral disease variants may be uncovered through evaluation of the underlying haplotypes. This methodology does not require the causal variant to be identified and tested directly, but rather has the potential to highlight physical regions that harbor putative disease-associated variants. In the present study, we have employed a genetic haplotype approach to examine the contribution of common variation at the *CYP19* locus to breast cancer risk among African-American, Hawaiian, Japanese, Latina and white women in the Multiethnic Cohort Study (MEC). In this study, we first defined LD blocks and constructed genetic haplotypes across the *CYP19* locus in a multiethnic panel. A reduced set of haplotype tagging SNPs (htSNPs) was selected that allow high predictability of the haplotypes within each block, and we evaluated these haplotypes in relationship with breast cancer risk in a large nested case–control study within the MEC. We also evaluated the independent effects of known missense variants.

## RESULTS

### LD and haplotype structure of the *CYP19* locus in the multiethnic panel

We assembled a high-density SNP map across the *CYP19* locus to determine LD block and haplotype structure (Fig. 1A); 74 SNPs were selected using an iterative strategy (see Methods) and the average distance between SNPs across the 189.4 kb region was 2.6 kb. We determined the *CYP19* locus to contain five blocks of LD (Fig. 1B; see Methods for block partitioning criteria): block 1 (SNPs 4–22) covered 38 kb, spanning exons I.1, 2a and I.4; block 2 (SNPs 24–36) spanned 32 kb, and encompassed exons I.5, I.7 and I.f; block 3 (SNPs 37–43) covered 13 kb and was located between exons I.f and I.2; and block 4 (SNPs 44–66) covered 50 kb and spanned the entire coding region, exons/promoters I.6, I.3 and PII through 5.8 kb downstream of exon 10. Block 5 was well downstream of *CYP19* and was not analyzed (see Methods). The linkage disequilibrium plot for the multiethnic sample is provided in Fig. 2. With this high-density SNP map we were able to narrow the intervals between blocks. The distances between blocks 1 and 2, 2 and 3, and 3 and 4 were ~7, 6 and 4 kb, respectively (Fig. 1B). The LD pattern across the locus was similar among Hawaiians, Japanese, Latinas and whites. For African-Americans, the size of most blocks was modestly reduced (block 1, SNPs 4–21, 35 kb; block 2, SNPs 24–35, 30 kb; and block 4, SNPs 44–65, 45 kb) and consequently distances between blocks were slightly greater.

Within each block, we observed low haplotype diversity (Fig. 1B) and, further, the majority of common haplotypes (i.e. ≥5% frequency) were shared across multiple ethnic groups (Table 1). For block 1, we observed eight common haplotypes (1a–1h) that could be predicted by six htSNPs. Block 2 was represented by five common haplotypes (2a–2e) that we could distinguish by six htSNPs, and block 3 contained six haplotypes (3a–3f) which may be described by five htSNPs. The fourth block was the largest and contained 10 common haplotypes (4a–4j) that could be defined by eight htSNPs. Within block 1, five of the eight common haplotypes (63%) were observed in more than one ethnic group, five of five in block 2 (100%), three of six in block 3 (50%), and seven of 10 in block 4 (70%). As expected, African-Americans displayed greater haplotype diversity, and four htSNPs (SNPs 14, 40, 41 and 52) were required only to distinguish African-American specific haplotypes (24). For each ethnic group, the common haplotypes (≥5%) comprised 85–100% of the total predicted haplotype variation within a defined block, and the average $R_h^2$ (see Methods) to predict the common haplotypes in the multiethnic panel was 0.92 (range 0.72–1.00; Table 1).

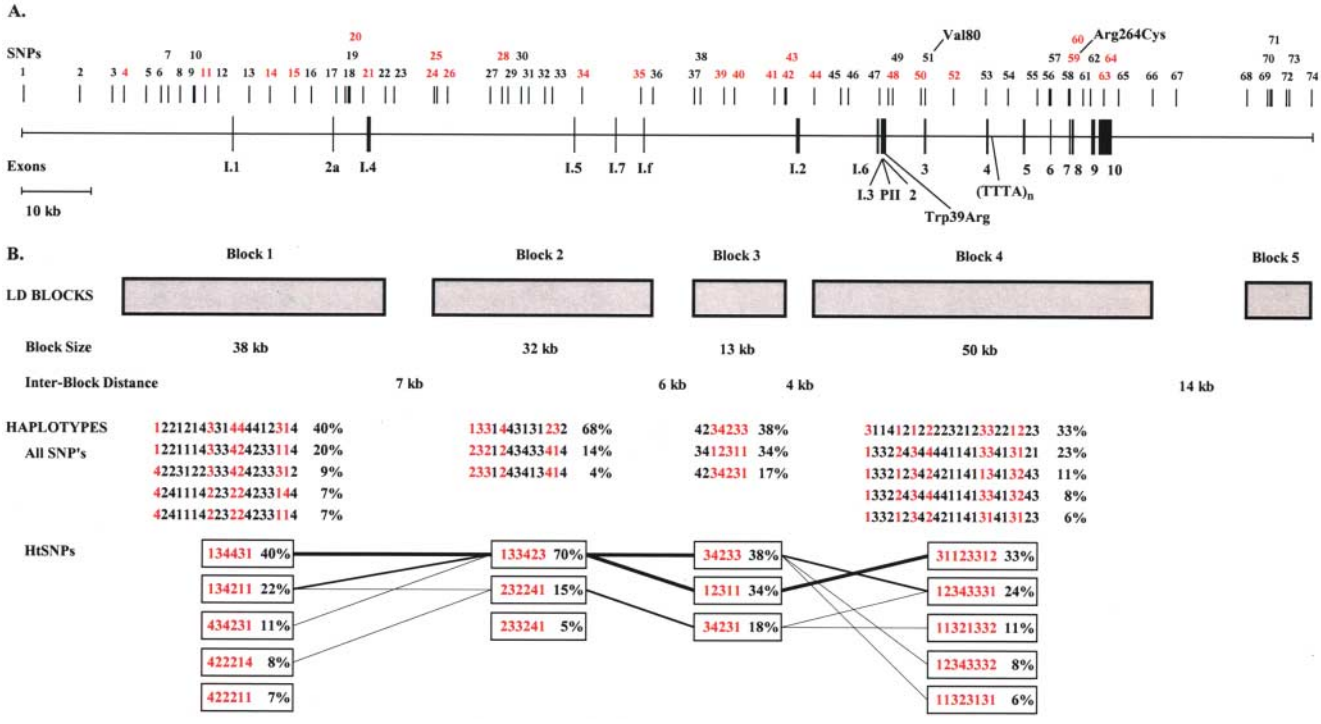


**Figure 1.** The genomic organization of *CYP19*. (**A**) The 74 SNPs used in the haplotype analysis. SNP location is based on the April 2003 freeze of chromosome 15 (contig NT_010194, http://genome.ucsc.edu). htSNPs for each block are indicated in red. (**B**) LD block and haplotype patterns across *CYP19*. Presented are the common haplotypes (≥5%) estimated using all SNPs and the htSNPs among all ethnic groups combined. The lines between blocks link haplotypes that are transmitted with ≥2.5% frequency across blocks. The numbers for each SNP correspond to the nucleotide at that position (1 = A, 2 = C, 3 = G, 4 = T).

### Breast cancer case–control analysis

Among all women, the mean age of the cases and controls was 64.3 and 63.4 years, respectively, and the mean age was similar for cases and controls within each ethnic group (Table 2). The distributions of established breast cancer risk factors were generally consistent with expectation, and were similar to what we observed in the overall cohort (26). Compared with controls, cases were more likely to be a current user of hormone replacement therapy and have a first-degree family history of breast cancer. Cases were also more likely to be nulliparous and to have had children at a later age. These associations were generally consistent across all ethnic groups.

The frequency of the common haplotypes (≥5%) predicted by the htSNPs in the multiethnic panel were nearly identical to those observed in the larger sample of cases and controls (Tables 1, 3 and 4). Three haplotypes that were observed at ≥5% frequency in at least one ethnic group in the multiethnic panel were <5% among cases and controls in each group and were not further evaluated in the case–control analysis (haplotypes 2e, 3f and 4j).

### Tests of haplotype associations

We first performed global tests for differences in risk according to the common haplotypes and observed marginally significant haplotype-effects in block 2 ($P = 0.01$), but not in blocks 1 ($P = 0.45$), 3 ($P = 0.14$) or 4 ($P = 0.45$). Within each block, we observed positive associations with individual haplotypes in an ethnic stratified analysis using all ethnic groups combined (Tables 3 and 4). In block 4, which spans the entire coding region of *CYP19*, we observed a non-significant positive association with haplotype 4c (OR = 1.13; 95% CI, 0.95–1.34; Table 3). In block 1, we observed a suggestive association with haplotype 1d (OR = 1.21; 95% CI, 1.02–1.43; Table 4), and in block 2, positive associations were noted with haplotypes 2b (OR = 1.23; 95% CI, 1.07–1.40) and 2d (OR = 1.28; 95% CI, 1.01–1.62). Within block 2, where the global test was statistically significant, the test for differences in risk between ethnic groups associated with haplotypes 2b and 2d was not significant ($P = 0.09$). In block 3, we also observed haplotype 3c to be associated with elevated risk (OR = 1.21; 95% CI, 1.05–1.39). When limiting the analysis to women with advanced disease (cases, $n = 342$) the associations with haplotypes 4c, 1d, 2d and 3c remained (data not shown), and for haplotype 2b, the strength of the association increased (OR = 1.41; 95% CI, 1.13–1.76).

### Evaluation of long-range haplotype patterns

Limited inter-block recombination may result in long-range LD, i.e. associations between haplotypes in adjacent blocks. In attempt to localize the signal in this region, we evaluated whether there was a long-range haplotype comprised of a subset of the common block-specific haplotypes (1d, 2b, 3c and 4c) that were associated with risk within each block. If a disease variant arose on a long-range haplotype then we would expect that the risk associated with this haplotype would be greater than the risk observed with each block-specific haplotype. The extent of coupling between haplotypes in

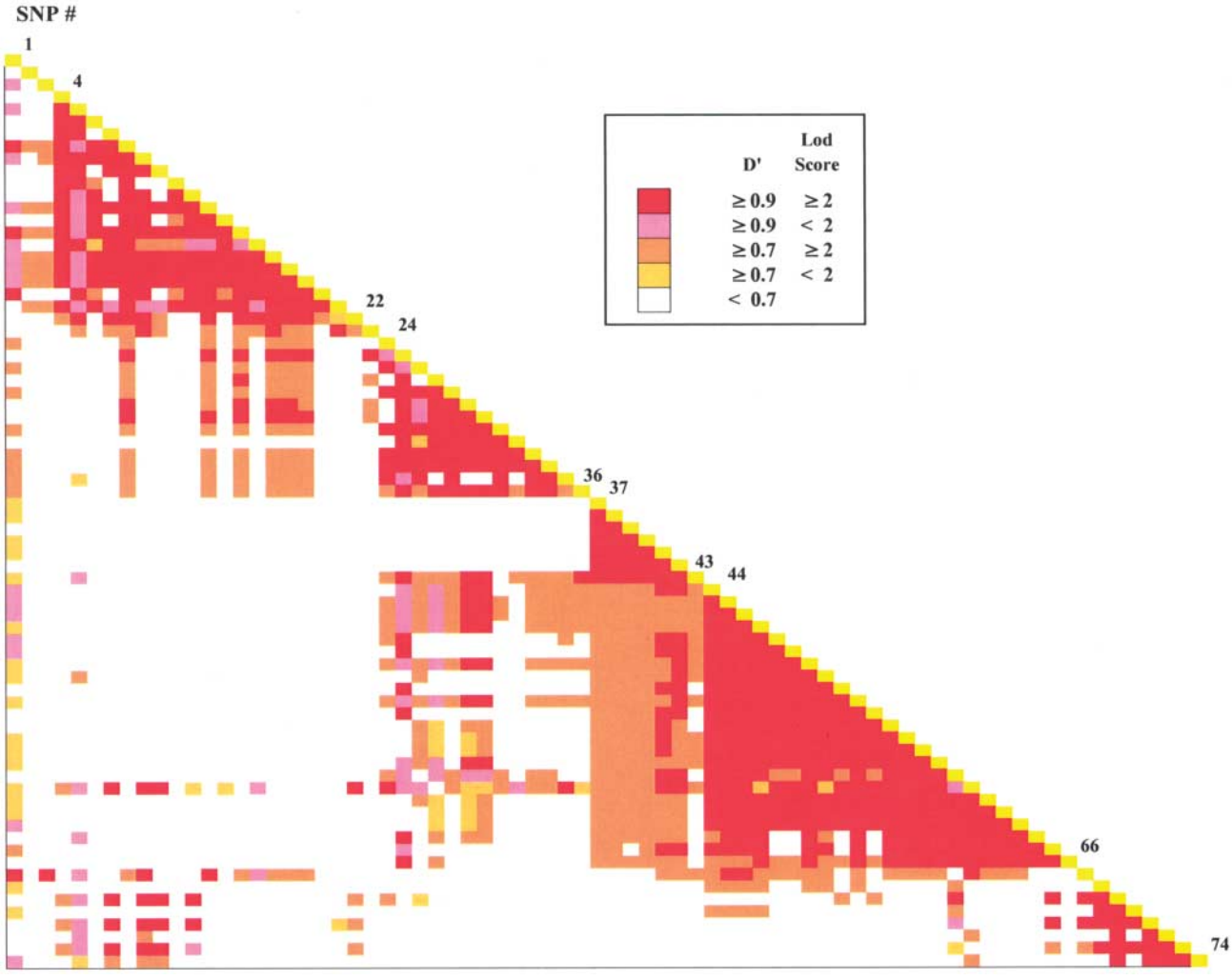


**Figure 2.** The linkage disequilibrium plot of *CYP19* for all ethnic groups combined. LD strength between the 74 SNPs, as indicated by the color scheme, was measured using a combination of the statistic $D'$ and LOD scores.

adjacent blocks varied considerably among the ethnic groups (Table 5). In analyses among all ethnic groups combined, we observed a strong association with the long-range haplotype 2b–3c (OR = 1.31; 95% CI, 1.11–1.54; Table 5) that was nominally greater than the associations observed with any block-specific haplotypes. The addition of haplotypes 1d and 4c to the long-range haplotype 2b–3c resulted in numerous less frequent haplotypes containing 2b–3c, especially among the African-Americans where 2b and 3c were more common, and did not increase the ORs. Therefore, the addition of 1d and 4c did not partition the 2b–3c haplotype in defining a common long-range haplotype that predicted greater risk beyond that observed with haplotype 2b–3c alone.

### Previously studied SNPs

Among the common haplotypes, the *Cys264* allele of the previously reported *Arg264Cys* polymorphism in exon 7 (27) is unique to haplotype 4c and, when evaluating this variant independently, we observed a modest non-significant positive association between the *Cys264* allele and breast cancer risk among all groups combined (OR = 1.18; 95% CI, 0.99–1.42; Table 6). The selection of this SNP as an htSNP in the multiethnic panel was not essential to define haplotype 4c, and thus we are unable to distinguish the effects of this SNP from haplotype 4c. We also evaluated two other well-studied SNPs in *CYP19*, the *Trp39Arg* missense variant in exon 2 (28,29) and a SNP in the 3′-UTR of exon 10 (30,31). The *Arg39* allele was only detected among the Hawaiians (2.1%) and the Japanese (2.9%) and was noted to travel exclusively on haplotype 4b. We observed little evidence of an association between the *Arg39* allele or the exon 10 variant and breast cancer risk (Table 6).

## DISCUSSION

In this study we have implemented an efficient stepwise approach that we are currently using to search for common disease alleles in candidate cancer susceptibility genes in the MEC. These studies are initiated by surveying variation across each gene in a multiethnic panel of subjects. This preliminary

**Table 1.** Common haplotypes in blocks 1–4 of *CYP19* among African-Americans, Hawaiians, Japanese, Latinas and whites in the multiethnic panel[a]

| Haplotypes[b] | Haplotypes frequencies in the multiethnic panel (%) | | | | |
|---|---|---|---|---|---|
| | African-Americans | Hawaiians | Japanese | Latinas | Whites |
| *Block 1 (SNPs 4–22) htSNPs: 4,11,14,15,20,21* | | | | | |
| 1a 134431 | 12 | 58 | 38 | 42 | 52 |
| 1b 134211 | 22 | 12 | 10 | 33 | 29 |
| 1c 434231 | | 12 | 32 | | |
| 1d 422214 | 9 | 5 | 12 | 7 | 7 |
| 1e 422211 | 15 | 7 | 5 | 5 | |
| 1f 134231 | 13 | | | | |
| 1g 432231 | 14 | | | | |
| 1h 424231 | | | | 6 | |
| Total[c] | 85 | 94 | 97 | 93 | 88 |
| $R_h^2$[d] | 0.85 | 0.91 | 0.93 | 0.97 | 0.89 |
| *Block 2 (SNPs 24–36) htSNPs: 24,25,26,28,34,35* | | | | | |
| 2a 133423 | 34 | 80 | 67 | 86 | 82 |
| 2b 232241 | 38 | 6 | 21 | 6 | 6 |
| 2c 233241 | 8 | 9 | 5 | | |
| 2d 143243 | 6 | | | 5 | 8 |
| 2e 233243 | 6 | | 6 | | |
| Total | 92 | 95 | 99 | 97 | 96 |
| $R_h^2$ | 0.85 | 0.73 | 0.96 | 1.00 | 0.72 |
| *Block 3 (SNPs 37–43) htSNPs: 39,40,41,42,43* | | | | | |
| 3a 34233 | 15 | 43 | 40 | 57 | 37 |
| 3b 12311 | 24 | 42 | 35 | 26 | 45 |
| 3c 34231 | 38 | 7 | 21 | 9 | 14 |
| 3d 12231 | 16 | | | | |
| 3e 12331 | | 8 | | | |
| 3f 14231 | 6 | | | | |
| Total | 99 | 100 | 96 | 92 | 96 |
| $R_h^2$ | 0.94 | 1.00 | 1.00 | 0.98 | 1.00 |
| *Block 4 (SNPs 44–66) htSNPs: 44,48,50,52,59,60,63,64* | | | | | |
| 4a 31123312 | 11 | 43 | 36 | 28 | 47 |
| 4b 12343331 | 14 | 35 | 28 | 23 | 19 |
| 4c 11321332 | 16 | | 25 | | |
| 4d 12343332 | | 6 | | 18 | 13 |
| 4e 11323131 | 6 | | | 16 | 9 |
| 4f 11323312 | 7 | 7 | | | |
| 4g 11323332 | 15 | | | | |
| 4h 11123312 | 5 | | | 5 | |
| 4i 11343332 | 9 | | | | |
| 4j 11323331 | 5 | | | | |
| Total | 88 | 91 | 89 | 90 | 88 |
| $R_h^2$ | 0.86 | 1.00 | 0.96 | 0.89 | 1.00 |

[a]Haplotypes observed with ≥5% frequency in at least on ethnic group in the multiethnic panel.
[b]Haplotype order is based on the frequency as predicted by the htSNPs among all groups combined.
[c]The percentage of all chromosomes accounted for by the common haplotypes.
[d]The $R_h^2$ that is given is the minimum $R_h^2$ of the common haplotypes in each ethnic group.

step allows us to determine which SNPs are polymorphic and assess allele frequencies in the different populations. This information is then used to establish the LD block structure, reconstruct haplotypes and select htSNPs that predict the common haplotypes in different ethnic populations.

Linkage disequilibrium blocks are regions that display little evidence of historical recombination and are characterized by low haplotype diversity (23–25). A previous study has demonstrated that, within LD blocks, more than 90% of the diversity of common haplotypes (>5%) may be captured by six to eight common SNPs (≥10%), and that these common haplotypes explain the vast majority of genetic variation contributed by unmeasured or undiscovered SNPs (24). We based our haplotype discovery and htSNP selection on this observation. In the present study, we selected a subset ($n = 74$) of all available SNPs (>250) from the private and public SNP databases to define LD blocks and reconstruct haplotypes across the *CYP19* gene. This process of defining LD blocks prior to haplotype estimation differs from other haplotype-based approaches where haplotypes are estimated without regard for the nature of LD across the candidate gene. The initial identification of LD blocks using a high-density set of available SNPs guarantees that common variation across each LD block is captured, which may not be the case when only a handful of SNPs are chosen based on convenience. In addition, picking htSNPs to predict the haplotypes within defined LD blocks results in a substantial reduction in genotyping required to study common variation across a candidate gene locus.

The LD block structure and haplotype diversity across *CYP19* was compatible with other studies that have explored more expansive regions of the human genome and consistent with prior observations of population differences in genetic diversity (24,32). In general, African-Americans were observed to have smaller LD blocks and a greater diversity of common haplotypes than the Hawaiians, Japanese, Latinas and whites. The ramifications for only using the available SNPs in the public and Celera databases in haplotype-based association studies in a multiethnic population are unclear. For example, it is estimated that only 80% of all common SNPs (>10%) among European Americans and 50% among African-Americans are in high correlation with SNPs in dbSNP, and it has been argued that the resequencing of candidate genes to uncover common ethnic-specific SNPs will be required for genetic LD association studies among African-Americans and other genetically diverse populations (33,34). Within LD blocks, however, a low haplotype diversity is observed which is not a consequence of genotyping only a subset of all available markers, but rather that recombination in the region is low so that SNPs are redundant in defining the common haplotypes. Using a high-density set of 74 common SNPs, spaced every 2.6 kb on average, we identified four LD blocks spanning the *CYP19* gene. To ensure adequate characterization of the common haplotypes, we obtained at least seven SNPs with frequencies ≥10% within an LD block. Within each block, more than 80% of the haplotype diversity could be accounted for by 5 to 10 common haplotypes and the majority of these haplotypes were observed to be shared across populations. It remains plausible that undiscovered ethnic-specific SNPs that are common may create subtypes of the major haplotype patterns which we have identified within the LD blocks. However, because we over-sampled SNPs within LD blocks, this is unlikely, and we feel confident that we were able to delineate the common haplotypes, especially within block 4, which spans the coding region, as we obtained 22 SNPs that were common among all ethnic groups.

In haplotype-based studies, the misclassification of rare haplotypes by grouping them with the more common haplotypes one is interested in evaluating may lead to the underestimation of haplotype-specific effects. In this study, we utilized

**Table 2.** Descriptive characteristics among breast cancer cases ($n = 1355$) and controls ($n = 2580$) in the Multiethnic Cohort Study

| | Ethnicity | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | African-Americans | | Hawaiians | | Japanese | | Latinas | | Whites | |
| | Cases ($n = 278$) | Controls ($n = 672$) | Cases ($n = 92$) | Controls ($n = 311$) | Cases ($n = 358$) | Controls ($n = 429$) | Cases ($n = 272$) | Controls ($n = 706$) | Cases ($n = 355$) | Controls ($n = 462$) |
| Age (mean) | 64.1 | 64.3 | 60.7 | 59.5 | 64.4 | 64.2 | 63.9 | 62.8 | 64.7 | 62.3 |
| Menopausal status (%) | | | | | | | | | | |
| Premenopausal | 14 | 11 | 17 | 27 | 13 | 21 | 10 | 11 | 8 | 20 |
| Postmenopausal[a] | 55 | 56 | 58 | 52 | 67 | 62 | 66 | 61 | 66 | 61 |
| Simple hysterectomy | 19 | 23 | 13 | 14 | 10 | 10 | 14 | 19 | 17 | 14 |
| Missing | 11 | 10 | 12 | 7 | 11 | 8 | 10 | 9 | 8 | 5 |
| HRT use (%)[a,b] | | | | | | | | | | |
| Never | 47 | 48 | 40 | 39 | 28 | 29 | 44 | 48 | 28 | 34 |
| Past | 29 | 23 | 15 | 23 | 16 | 15 | 21 | 19 | 17 | 17 |
| Current | 21 | 26 | 43 | 36 | 56 | 53 | 31 | 28 | 55 | 48 |
| Age at menarche (%)[b] | | | | | | | | | | |
| ≤12 | 54 | 46 | 57 | 59 | 56 | 50 | 48 | 49 | 55 | 48 |
| 13–14 | 34 | 39 | 28 | 29 | 32 | 35 | 37 | 39 | 35 | 44 |
| 15+ | 12 | 14 | 11 | 11 | 9 | 14 | 13 | 11 | 8 | 8 |
| Number of children (%)[b] | | | | | | | | | | |
| 0 | 12 | 11 | 9 | 9 | 16 | 11 | 10 | 7 | 18 | 16 |
| 1 | 20 | 16 | 3 | 10 | 11 | 10 | 8 | 6 | 11 | 9 |
| 2 or 3 | 40 | 40 | 48 | 39 | 54 | 60 | 36 | 36 | 51 | 53 |
| 4+ | 25 | 32 | 40 | 42 | 17 | 18 | 45 | 50 | 19 | 22 |
| Age at first birth (%)[b,c] | | | | | | | | | | |
| <20 | 46 | 50 | 43 | 39 | 8 | 11 | 35 | 41 | 23 | 23 |
| 21–30 | 44 | 41 | 52 | 50 | 74 | 75 | 53 | 52 | 66 | 63 |
| 31+ | 8 | 5 | 0 | 7 | 14 | 12 | 9 | 4 | 10 | 12 |
| First degree family history of breast cancer (%)[b] | | | | | | | | | | |
| Yes | 23 | 12 | 15 | 14 | 18 | 11 | 17 | 10 | 15 | 9 |
| No | 72 | 82 | 80 | 82 | 77 | 87 | 75 | 83 | 81 | 88 |
| Average alcohol consumption (drinks/day)[b] | | | | | | | | | | |
| 0 | 51 | 53 | 66 | 57 | 72 | 75 | 54 | 53 | 31 | 39 |
| <1 | 29 | 30 | 17 | 30 | 19 | 17 | 32 | 33 | 39 | 40 |
| ≥1 | 10 | 10 | 11 | 9 | 5 | 3 | 5 | 6 | 22 | 18 |

[a]Women reporting natural menopause or having had a bilateral oophorectomy.
[b]Numbers do not add to 100% because of missing data.
[c]Among parous women.

a formal measure, $R_h^2$, to select the htSNPs for predicting the common haplotypes (35). This approach optimizes one's ability to identify a specific haplotype and not merely distinguish different clades (groups of related haplotypes). This high degree of predictability reduces the potential bias incurred from haplotype misclassification. In this study, the average $R_h^2$ for defining the common haplotypes was 0.92. With our sample size and assuming a dominant inheritance model, we had more than 90% power to detect relative risks as low as 1.27 for a common haplotype (25% frequency) that was shared across populations.

Rare mutations in *CYP19* that result in substantial reductions in enzyme activity have been reported in patients with aromatase deficiency (36). More common variation in *CYP19* has been hypothesized to contribute to phenotypes associated with estrogen and androgen exposure such as breast and prostate cancer (37). In this study, we observed modest associations between block-specific haplotypes of *CYP19* and increased breast cancer risk. These associations were not observed consistently across the groups, although we had limited power to detect ethnic-specific risks because of low haplotype frequencies in some groups. We studied the long-range haplotype patterns across LD blocks in an attempted to localize the region containing a putative disease variant. Our data suggest that a susceptibility allele may have arisen on a particular long-range haplotype that contains haplotypes 2b and 3c, but additional studies will be required to confirm these findings before undertaking a resequencing of this region among individuals with this haplotype combination.

Previous association studies have focused on a limited set of polymorphisms at the *CYP19* locus. The most well-studied polymorphism in *CYP19* has been the tetranucleotide $(TTTA)_n$ repeat in intron 4, and positive associations have been noted for the rare 10 and 12 repeat alleles (18,21). Based on the location of this repeat polymorphism, it is not likely to be functional, and if it is a marker of risk it is probably because it is in LD with functional variants elsewhere in the gene. Further work will be required to determine whether these repeat alleles mark the haplotypes that we observed to be associated with greater risk. Studies in Asian populations have provided little support for

**Table 3.** Associations between haplotypes in LD block 4 of *CYP19* and breast cancer risk

| Haplotypes[a] Block 4 | Haplotype frequencies | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | African-Americans | | | Hawaiians | | | Japanese | | |
| | Cases ($n=266$) | Controls ($n=651$) | OR[b] (95% CI) | Cases ($n=78$) | Controls ($n=295$) | OR[b] (95% CI) | Cases ($n=347$) | Controls ($n=420$) | OR[b] (95% CI) |
| 4a 31123312 | 13.2 | 12.9 | Ref | 42.9 | 40.4 | Ref | 31.3 | 36.4 | Ref |
| 4b 12343331 | 20.3 | 16.9 | 1.17 (0.80–1.69) | 28.9 | 35.2 | 0.78 (0.50–1.21) | 23.1 | 24.8 | 1.05 (0.81–1.37) |
| 4c 11321332 | 15.1 | 14.3 | 1.03 (0.70–1.51) | 6.4 | 3.6 | 1.70 (0.77–3.76) | 29.7 | 25.7 | 1.31 (1.02–1.68) |
| 4d 12343332 | | | | 6.4 | 4.2 | 1.52 (0.66–3.47) | | | |
| 4e 11323131 | 6.4 | 6.8 | 0.91 (0.56–1.49) | | | | | | |
| 4f 11323312 | 4.7 | 5.5 | 0.84 (0.49–1.43) | 5.7 | 6.3 | 0.80 (0.36–1.77) | 8.6 | 6.6 | 1.46 (0.97–2.20) |
| 4g 11323332 | 12.5 | 12.8 | 0.96 (0.64–1.44) | | | | | | |
| 4h 11123312 | 5.5 | 6.8 | 0.79 (0.47–1.33) | | | | | | |
| 4i 11343332 | 8.9 | 9.0 | 0.96 (0.61–1.50) | | | | | | |
| 4j 11323331 | | | | | | | | | |

| Haplotypes[a] Block 4 | Haplotype frequencies | | | | | | |
|---|---|---|---|---|---|---|---|
| | Latinas | | | Whites | | | All groups combined |
| | Cases ($n=254$) | Controls ($n=673$) | OR[b] (95% CI) | Cases ($n=342$) | Controls ($n=443$) | OR[b] (95% CI) | OR[c] (95% CI) |
| 4a 31123312 | 24.4 | 26.2 | Ref | 44.5 | 39.9 | Ref | Ref |
| 4b 12343331 | 29.0 | 22.5 | 1.35 (1.02–1.79) | 17.2 | 18.8 | 0.79 (0.59–1.05) | 1.05 (0.91–1.20) |
| 4c 11321332 | 5.3 | 4.3 | 1.25 (0.75–2.07) | | | | 1.13 (0.95–1.34) |
| 4d 12343332 | 11.5 | 14.4 | 0.83 (0.57–1.20) | 15.9 | 14.7 | 0.98 (0.72–1.32) | 0.97 (0.79–1.19) |
| 4e 11323131 | 14.9 | 18.5 | 0.88 (0.64–1.21) | 7.9 | 10.3 | 0.71 (0.49–1.03) | 0.81 (0.66–0.99) |
| 4f 11323312 | | | | | | | 1.07 (0.84–1.36) |
| 4g 11323332 | | | | | | | 0.99 (0.74–1.33) |
| 4h 11123312 | 3.9 | 5.3 | 0.85 (0.50–1.47) | 6.7 | 5.3 | 1.15 (0.73–1.80) | 0.95 (0.73–1.23) |
| 4i 11343332 | | | | | | | 1.03 (0.73–1.44) |
| 4j 11323331 | | | | | | | |

[a]Haplotypes observed with ≥5% frequency among cases or controls in at least on ethnic group are shown.
[b]ORs are estimated using unconditional logistic regression adjusted for age.
[c]ORs are estimated using unconditional logistic regression adjusted for age and ethnicity.

the *Cys264* allele as a breast cancer risk factor (28,38,39). In our study, the *Cys264* allele was more common among the Japanese and African-Americans (>14%) and was only modestly associated with increased risk. Combining the data from the previous three studies, the OR for the *Cys264* allele is OR = 1.03 (95% CI, 0.83–1.28) and the 95% confidence interval is compatible with the effect we observed. In addition, our findings do not support previous reports suggesting that carriers of the *Arg39* allele are at lower risk of breast cancer (28,29).

A strength of the present study is the large sample size among each of five ethnic populations. This study design enables the reproducibility of an association to be evaluated across multiple ethnic groups, providing more convincing support for an underlying relationship between a genetic marker and breast cancer risk. Our sample size within each ethnic group however, is not large enough to definitively evaluate ethnic-specific risks. In addition, our findings must be interpreted with caution as numerous statistical tests were conducted separately for multiple block-specific haplotypes and long-range haplotype combinations.

This comprehensive genetic analysis provides a framework for haplotype-based studies of *CYP19* in relationship with other phenotypes for which steroid hormones have been implicated, such as stature, obesity and diabetes (40). Although these data provide little support for there being a strong breast cancer susceptibility allele at the *CYP19* locus that is common in the general population, they do suggest that individuals with the long-range haplotype 2b–3c may harbor a variant that modestly increases risk.

## MATERIALS AND METHODS

### The Multiethnic Cohort

The MEC consists of over 215 000 men and women in Hawaii and Los Angeles (with additional African-Americans from elsewhere in California) and has been described in detail elsewhere (41). In brief, the cohort is comprised predominantly of Hawaiians, Japanese and whites in Hawaii, and African-Americans, Japanese and Latinos in Los Angeles. Between 1993 and 1996, participants entered the MEC by completing a 26-page self-administered mail questionnaire that asked detailed information about dietary habits, demographic factors (ethnicity, education and migrant status), personal behaviors (smoking, sun exposure and physical activity), history of prior medical conditions (e.g. heart attack, diabetes and cancer), family history of common cancers, and for women, reproductive history and exogenous hormone use. Potential cohort members were identified through the Department of Motor Vehicles drivers' license files, and additionally for African-Americans, Health Care Financing Administration data files. The participants were between the ages 45 and 75 when they entered the cohort.

**Table 4.** Associations between haplotypes in LD blocks 1–3 of *CYP19* and breast cancer risk

| Haplotypes[a] Block 1 | Haplotype frequencies African-Americans | | | Hawaiians | | | Japanese | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cases ($n=266$) | Controls ($n=651$) | OR[b] (95% CI) | Cases ($n=78$) | Controls ($n=295$) | OR[b] (95% CI) | Cases ($n=347$) | Controls ($n=420$) | OR[b] (95% CI) |
| 1a 134431 | 17.3 | 17.1 | Ref | 51.6 | 56.1 | Ref | 34.6 | 39.0 | Ref |
| 1b 134211 | 20.1 | 20.6 | 0.98 (0.70–1.36) | 8.3 | 10.8 | 0.77 (0.40–1.51) | 14.8 | 12.8 | 1.29 (0.94–1.79) |
| 1c 434231 | 5.8 | 4.6 | 1.30 (0.77–2.20) | 16.1 | 15.6 | 1.10 (0.67–1.81) | 19.5 | 22.2 | 0.97 (0.74–1.29) |
| 1d 422214 | 4.9 | 6.3 | 0.78 (0.48–1.27) | 10.9 | 7.6 | 1.66 (0.91–3.02) | 19.4 | 15.3 | 1.44 (1.07–1.93) |
| 1e 422211 | 12.2 | 12.9 | 0.95 (0.65–1.38) | 7.0 | 5.1 | 1.77 (0.81–3.89) | 5.9 | 5.8 | 1.13 (0.72–1.76) |
| 1f 134231 | 11.8 | 11.5 | 1.00 (0.67–1.49) | | | | | | |
| 1g 432231 | 15.8 | 15.0 | 1.04 (0.73–1.49) | | | | | | |
| 1h 424231 | | | | | | | | | |

| Haplotypes[a] Block 1 | Haplotype frequencies Latinas | | | Whites | | | All groups combined |
|---|---|---|---|---|---|---|---|
| | Cases ($n=254$) | Controls ($n=673$) | OR[b] (95% CI) | Cases ($n=342$) | Controls ($n=443$) | OR[b] (95% CI) | OR[c] (95% CI) |
| 1a 134431 | 42.1 | 37.5 | Ref | 47.2 | 49.7 | Ref | Ref |
| 1b 134211 | 28.3 | 31.9 | 0.81 (0.63–1.04) | 29.2 | 27.6 | 1.14 (0.89–1.46) | 1.02 (0.89–1.16) |
| 1c 434231 | | | | | | | 1.00 (0.83–1.22) |
| 1d 422214 | 12.5 | 10.4 | 1.09 (0.78–1.52) | 9.6 | 8.3 | 1.24 (0.86–1.78) | 1.21 (1.02–1.43) |
| 1e 422211 | 3.1 | 5.3 | 0.54 (0.30–0.95) | | | | 0.93 (0.75–1.15) |
| 1f 134231 | | | | | | | 0.92 (0.70–1.19) |
| 1g 432231 | | | | | | | 1.08 (0.82–1.42) |
| 1h 424231 | 5.5 | 5.9 | 0.84 (0.53–1.35) | 4.7 | 5.6 | 0.90 (0.56–1.45) | 1.15 (0.88–1.49) |

| Haplotypes[a] Block 2 | Haplotype frequencies African-Americans | | | Hawaiians | | | Japanese | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cases ($n=266$) | Controls ($n=651$) | OR[b] (95% CI) | Cases ($n=78$) | Controls ($n=295$) | OR[b] (95% CI) | Cases ($n=347$) | Controls ($n=420$) | OR[b] (95% CI) |
| 2a 133423 | 40.2 | 41.0 | Ref | 72.4 | 77.6 | Ref | 53.9 | 61.3 | Ref |
| 2b 232241 | 36.6 | 31.8 | 1.20 (0.95–1.51) | 11.0 | 10.0 | 1.17 (0.65–2.09) | 29.5 | 23.3 | 1.42 (1.13–1.80) |
| 2c 233241 | 3.9 | 7.0 | 0.56 (0.34–0.94) | 6.9 | 7.1 | 0.96 (0.46–2.00) | 13.2 | 10.5 | 1.43 (1.03–1.98) |
| 2d 143243 | 9.8 | 7.9 | 1.29 (0.89–1.87) | | | | | | |
| 2e 233243 | | | | | | | | | |

| Haplotypes[a] Block 2 | Haplotype frequencies Latinas | | | Whites | | | All groups combined |
|---|---|---|---|---|---|---|---|
| | Cases ($n=254$) | Controls ($n=673$) | OR[b] (95% CI) | Cases ($n=342$) | Controls ($n=443$) | OR[b] (95% CI) | OR[c] (95% CI) |
| 2a 133423 | 78.6 | 83.5 | Ref | 85.5 | 83.3 | Ref | Ref |
| 2b 232241 | 11.1 | 8.9 | 1.29 (0.93–1.79) | 3.6 | 5.9 | 0.58 (0.35–0.96) | 1.23 (1.07–1.40) |
| 2c 233241 | | | | | | | 1.06 (0.84–1.34) |
| 2d 143243 | 5.0 | 3.3 | 1.66 (0.98–2.79) | 6.2 | 6.0 | 1.02 (0.66–1.58) | 1.28 (1.01–1.62) |
| 2e 233243 | | | | | | | |

| Haplotypes[a] Block 3 | Haplotype frequencies African-Americans | | | Hawaiians | | | Japanese | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cases ($n=266$) | Controls ($n=651$) | OR[b] (95% CI) | Cases ($n=78$) | Controls ($n=295$) | OR[b] (95% CI) | Cases ($n=347$) | Controls ($n=420$) | OR[b] (95% CI) |
| 3a 34233 | 20.5 | 20.9 | Ref | 38.3 | 42.1 | Ref | 29.9 | 32.5 | Ref |
| 3b 12311 | 26.4 | 26.0 | 1.04 (0.78–1.41) | 43.5 | 42.1 | 1.14 (0.76–1.69) | 26.8 | 32.3 | 0.90 (0.72–1.19) |
| 3c 34231 | 35.6 | 32.1 | 1.14 (0.86–1.51) | 10.9 | 7.8 | 1.59 (0.82–3.05) | 30.0 | 23.5 | 1.40 (1.07–1.83) |
| 3d 12231 | 12.2 | 14.7 | 0.85 (0.59–1.21) | | | | | | |
| 3e 12331 | | | | 5.2 | 6.4 | 0.90 (0.40–2.06) | 7.7 | 6.4 | 1.35 (0.88–2.06) |
| 3f 14231 | | | | | | | | | |

**Table 4.** (*Continued*)

| Haplotypes[a] Block 3 | Haplotype frequencies Latinas Cases ($n=254$) | Controls ($n=673$) | OR[b] (95% CI) | Whites Cases ($n=342$) | Controls ($n=443$) | OR[b] (95% CI) | All groups combined OR[c] (95% CI) |
|---|---|---|---|---|---|---|---|
| 3a 34233 | 51.4 | 54.9 | Ref | 41.7 | 43.3 | Ref | Ref |
| 3b 12311 | 27.0 | 28.1 | 1.02 (0.80–1.30) | 43.8 | 38.8 | 1.19 (0.95–1.49) | 1.06 (0.94–1.19) |
| 3c 34231 | 13.2 | 10.5 | 1.34 (0.96–1.86) | 11.5 | 15.0 | 0.81 (0.59–1.11) | 1.21 (1.05–1.39) |
| 3d 12231 | | | | | | | 0.95 (0.71–1.27) |
| 3e 12331 | | | | | | | 1.11 (0.83–1.49) |
| 3f 14231 | | | | | | | |

[a]Haplotypes observed with ≥5% frequency among cases or controls in at least on ethnic group are shown.
[b]ORs are estimated using unconditional logistic regression adjusted for age.
[c]ORs are estimated using unconditional logistic regression adjusted for age and ethnicity.

**Table 5.** Associations of long-range *CYP19* haplotypes and breast cancer risk

| Haplotypes[a] | | | | Haplotype frequencies African-Americans | | Hawaiians | | Japanese | | Latinas | | Whites | | All groups combined OR (95% CI)[b] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Block 1 | 2 | 3 | 4 | Cases | Controls | Cases | Controls | Cases | Controls | Cases | Controls | Cases | Controls | |
| 1d | | | | 4.9 | 6.3 | 10.9 | 7.6 | 19.4 | 15.3 | 12.5 | 10.4 | 9.6 | 8.3 | 1.21 (1.02–1.43) |
| | 2b | | | 36.6 | 31.8 | 11.0 | 10.0 | 29.5 | 23.3 | 11.1 | 8.9 | 3.6 | 5.9 | 1.23 (1.07–1.40) |
| | | 3c | | 35.6 | 32.1 | 10.9 | 7.8 | 30.0 | 23.5 | 13.2 | 10.5 | 11.5 | 15.0 | 1.21 (1.05–1.39) |
| | | | 4c | 15.1 | 14.3 | 6.4 | 3.6 | 29.7 | 25.7 | 5.3 | 4.3 | | | 1.13 (0.95–1.34) |
| 1d | 2b | | | 3.3 | 3.8 | 4.1 | 3.9 | 16.6 | 13.0 | 5.5 | 3.9 | | | 1.23 (0.99–1.55) |
| | 2b | 3c | | 27.7 | 23.3 | 7.4 | 4.7 | 28.9 | 22.3 | 6.1 | 4.7 | 2.5 | 4.2 | 1.31 (1.11–1.54) |
| | | 3c | 4c | 12.5 | 10.2 | 5.7 | 2.9 | 26.4 | 21.0 | 4.5 | 3.8 | 2.2 | 3.7 | 1.26 (1.05–1.52) |
| | 2b | 3c | 4c | 11.3 | 9.3 | 5.8 | 2.6 | 26.1 | 20.5 | 4.3 | 3.6 | 2.2 | 3.5 | 1.31 (1.08–1.58) |
| 1d | 2b | 3c | | 3.1 | 3.2 | 3.7 | 3.1 | 16.2 | 12.3 | | | | | 1.28 (0.97–1.69) |

[a]Haplotypes observed with ≥5% frequency among cases or controls in at least one ethnic group (haplotypes ≥2.5% among cases or controls are shown).
[b]ORs are estimated using unconditional logistic regression adjusted for age and ethnicity.

Incident cancers in the MEC are identified by cohort linkage to population-based cancer Surveillence, Epidemiology and End Results (SEER) registries covering Hawaii and Los Angeles County, and to the California State cancer registry covering all of California. Case ascertainment in the SEER program is 98% (http://seer.cancer.gov/about/quality.html). Information on stage of disease at the time of diagnosis is also collected from the cancer registries. Women were classified as having advanced, high stage disease if they had non-localized breast cancer.

Beginning in 1994, blood samples were collected from incident breast cancer cases. At this time, blood collection was also initiated in a random sample of MEC participants to serve as a control pool for genetic analyses in the cohort. The participation rates for providing a blood sample were 74 and 66% for cases and controls, respectively; the difference in participation rates between cases with high and low stage disease was <10%. Eligible cases in this nested breast cancer case–control study consisted of women with incident breast cancer (including second primaries) diagnosed after enrollment in the MEC through May 2002. Controls were women without breast cancer prior to entry into the cohort and without a diagnosis up to May 2002. The breast cancer case–control study consists of 1355 breast cancer cases and 2580 controls. This study was approved by the Institutional Review Boards at the University of Southern California and at the University of Hawaii.

## SNP selection and genotyping in the multiethnic panel

We surveyed genetic variation across 189.4 kb spanning the *CYP19* locus, from 30.8 kb upstream of exon I.1 (the furthest 5′ first exon) through 29.4 kb downstream of the transcribed region. We attempted to select SNPs every 3–5 kb across the locus to ensure a high density of markers of moderate allele frequency and to provide adequate characterization of genetic haplotype diversity within defined LD blocks. SNPs were selected in an iterative manner and added until we had six to eight common SNPs (≥10%) per LD block and the distance between adjacent blocks was <10 kb. We included all known SNPs in the coding region. We selected 73 SNPs from the National Center for Biotechnology Information SNP database (www.ncbi.nlm.nih.gov/SNP/), 28 from the Celera database (www.celera.com) and two from the literature (19). SNPs were genotyped in a sample of 349 women in the MEC without a history of cancer: African-American ($n=70$), Hawaiian ($n=69$), Japanese ($n=70$), Latina ($n=70$) and

**Table 6.** Associations between SNPs in *CYP19* and breast cancer risk

| | African-Americans | Hawaiians | Japanese | Latinas | Whites | All groups[b] |
|---|---|---|---|---|---|---|
| *Trp39Arg* (rs2236722) | | | | | | |
| *Arg* allele frequency among controls | | 2.1 | 2.9 | | | |
| *Arg*/*Trp* versus *Trp*/*Trp* genotypes | | | | | | |
| OR (95% CI)[a] | | 0.39 (0.05–3.06) | 1.39 (0.79–2.46) | | | 1.34 (0.81–2.24) |
| *Arg264Cys*: SNP59 (rs700519) | | | | | | |
| *Cys* allele frequency among controls | 14.8 | 3.4 | 26.7 | 4.5 | 4.1 | |
| *Cys/Cys* + *Cys/Arg* versus *Arg/Arg* | | | | | | |
| OR (95% CI)[a] | 1.09 (0.80–1.50) | 2.40 (1.04–5.49) | 1.42 (1.06–1.89) | 1.19 (0.73–-1.92) | 0.52 (0.28–0.97) | 1.18 (0.99–1.42) |
| Exon 10, 3′-UTR: SNP63 (rs10046) | | | | | | |
| *T* allele frequency among controls | 26.2 | 49.6 | 44.1 | 34.9 | 47.8 | |
| *TC* versus *CC* OR (95% CI)[a] | 0.90 (0.66–1.22) | 1.39 (0.72–2.68) | 0.87 (0.63–1.21) | 0.81 (0.59–1.10) | 1.18 (0.82–1.70) | 0.93 (0.79–1.08) |
| *TT* versus *CC* OR (95% CI)[a] | 0.80 (0.44–1.45) | 1.17 (0.54–2.52) | 0.83 (0.55–1.23) | 0.95 (0.60–1.50) | 1.70 (1.12–2.59) | 1.04 (0.84–1.28) |
| *P*-trend | 0.35 | 0.71 | 0.32 | 0.45 | 0.01 | |

[a]Adjusted for age.
[b]Adjusted for age and ethnicity.

white ($n = 70$). This sample size guaranteed that any haplotype with a frequency of ≥5% will be represented at least once among the 140 chromosomes with probability >99%. The following SNPs were removed from the haplotype analysis: eight that were monomorphic or had minor allele frequencies <5% in all ethnic groups, 16 assays that provided poor genotyping results and five SNPs that appeared to have been mis-mapped during genome assembly based on LD relationships with other SNPs, leaving 74 SNPs with minor allele frequencies ≥5% in at least one ethnic group to include in the haplotype analysis (Fig. 1A, Table 7). We tested for Hardy–Weinberg equilibrium using the $\chi^2$ test with 1 d.f.; the observed genotype distributions based on allele frequencies for all 74 SNPs were consistent with Hardy–Weinberg equilibrium in at least four of the five ethnic groups. Two SNPs, 72 and 74, were not in Hardy–Weinberg equilibrium among the Japanese. These SNPs were not included in the haplotype analysis because they were located in block 5, which was not evaluated (see below).

DNA for the multiethnic panel was extracted from white blood cell fractions using the Qiagen Blood Kit (Qiagen, Chatsworth, CA, USA). Genotyping was performed by time-of-flight mass spectrometry (MALDI-TOF) using the Sequenom platform at the Whitehead Institute/MIT Center for Genome Research. Replicate blinded quality control samples (10%) were included to assess reproducibility of the genotyping procedure; less than 0.2% (4/2625) of the matched quality control pairs were discordant.

## Haplotype block determination

The $D'$ statistic was used as a pair-wise measure of linkage disequilibrium between the 74 SNPs used in the haplotype analysis (42). LD block structure was examined using the criteria of Gabriel *et al.* (24), which utilizes the 90% confidence bounds of $D'$ to define sites of historical recombination between SNPs (24). Block structure was assessed using SNPs with minor allele frequencies ≥10%. Blocks were initially defined following alignment across ethnic groups; borders were characterized by SNPs at the extreme ends of the block in any one ethnic group, except for African-Americans, whose block sizes (extent of LD), as expected, were modestly smaller than the other groups. We tested the suitability of this block definition by evaluating whether SNPs surrounding presumed block borders modified the number or identity of common haplotypes estimated within the blocks; changes in the number of haplotypes and the introduction of recombinant haplotypes would indicate whether SNPs were spanning a potentially important site of historical recombination and guided us in redefining a block boundary. We included SNPs with minor allele frequencies as low as 5% to both extend block boundaries defined using the criteria of Gabriel *et al.* (24) as well as to fully describe the diversity of the underlying common haplotypes in each ethnic group. Based on this information we determined that the *CYP19* locus could be parsed into four or five haplotype blocks depending on ethnicity. For African-Americans and Latinos, a clear site of recombination was observed between SNPs 43 and 44 that was not as evident among Hawaiians, Japanese or whites. This resulted in two distinct blocks (3 and 4) that were evaluated independently for all groups. The shared LD block structure between African-Americans and Latinas most likely reflects the recent admixture between these populations. Block 5 was located 14 kb from block 4 and greater than 20 kb 3′ of the transcribed region. This fifth block is less likely to contain a variant relevant to *CYP19* and was not further examined in the case–control haplotype analysis.

**Table 7.** Seventy-four SNPs used in the haplotype analysis of *CYP19*

| SNP no. | SNP ID | Position[a] | Minor allele | Minor allele frequency | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | African-Americans | Hawaiians | Japanese | Latinas | Whites |
| 1 | rs764531 | 49240734 | T | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | rs2445781 | 49232398 | G | 0.42 | 0.09 | 0.08 | 0.40 | 0.22 |
| 3 | rs2124874 | 49227735 | A | 0.09 | 0.25 | 0.56 | 0.21 | 0.19 |
| 4 | rs2446405 | 49225931 | A | 0.47 | 0.72 | 0.50 | 0.76 | 0.87 |
| 5 | rs2470162 | 49222705 | G | 0.02 | 0.02 | 0.02 | 0.05 | 0.03 |
| 6 | rs1551656 | 49220587 | T | 0.49 | 0.13 | 0.19 | 0.17 | 0.08 |
| 7 | rs2445771 | 49219497 | G | 0.29 | 0.15 | 0.33 | 0.09 | 0.05 |
| 8 | rs2470164 | 49217699 | C | 0.12 | 0.62 | 0.39 | 0.42 | 0.58 |
| 9 | rs1870050 | 49215689 | C | 0.04 | 0.15 | 0.33 | 0.09 | 0.05 |
| 10 | rs868475 | 49215592 | C | 0.25 | 0.14 | 0.33 | 0.09 | 0.05 |
| 11 | rs2445765 | 49214036 | C | 0.29 | 0.11 | 0.16 | 0.18 | 0.10 |
| 12 | rs1071955 | 49511950 | C | 0.29 | 0.10 | 0.18 | 0.12 | 0.08 |
| 13 | rs2446410 | 49207583 | A | 0.10 | 0.62 | 0.39 | 0.43 | 0.57 |
| 14 | rs1870049 | 49204361 | C | 0.44 | 0.13 | 0.18 | 0.15 | 0.10 |
| 15 | rs2470144 | 49200863 | T | 0.12 | 0.58 | 0.38 | 0.42 | 0.51 |
| 16 | rs2470145 | 49198276 | C | 0.05 | 0.02 | 0.01 | 0.07 | 0.04 |
| 17 | rs2445761 | 49194754 | T | 0.14 | 0.59 | 0.38 | 0.43 | 0.56 |
| 18 | rs2470147 | 49193301 | A | 0.12 | 0.60 | 0.38 | 0.41 | 0.56 |
| 19 | rs1902585 | 49193044 | C | 0.13 | 0.60 | 0.38 | 0.41 | 0.55 |
| 20 | rs1004984 | 49192667 | G | 0.46 | 0.76 | 0.72 | 0.51 | 0.61 |
| 21 | rs1902584 | 49190792 | T | 0.09 | 0.05 | 0.12 | 0.07 | 0.07 |
| 22 | rs1902583 | 49187589 | C | 0.38 | 0.15 | 0.34 | 0.10 | 0.06 |
| 23 | rs2470151 | 49186207 | C | 0.48 | 0.77 | 0.53 | 0.78 | 0.75 |
| 24 | hCV1664178 | 49180279 | A | 0.44 | 0.83 | 0.68 | 0.91 | 0.93 |
| 25 | rs2445759 | 49179979 | T | 0.05 | 0.02 | 0.00 | 0.05 | 0.08 |
| 26 | hCV1664175 | 49178491 | C | 0.41 | 0.09 | 0.21 | 0.06 | 0.05 |
| 27 | rs2470153 | 49172055 | G | 0.09 | 0.03 | 0.00 | 0.05 | 0.11 |
| 28 | rs730154 | 49170342 | T | 0.36 | 0.80 | 0.68 | 0.86 | 0.82 |
| 29 | hCV3060059 | 49169551 | C | 0.14 | 0.03 | 0.00 | 0.05 | 0.11 |
| 30 | rs2470158 | 49167533 | A | 0.15 | 0.04 | 0.00 | 0.05 | 0.11 |
| 31 | rs936309 | 49166517 | A | 0.35 | 0.79 | 0.34 | 0.85 | 0.78 |
| 32 | rs2470177 | 49164123 | A | 0.19 | 0.12 | 0.12 | 0.07 | 0.12 |
| 33 | rs2470176 | 49163077 | A | 0.38 | 0.81 | 0.68 | 0.86 | 0.82 |
| 34 | rs936306 | 49158736 | C | 0.36 | 0.80 | 0.68 | 0.86 | 0.83 |
| 35 | hCV11484670 | 49149991 | G | 0.49 | 0.85 | 0.74 | 0.91 | 0.94 |
| 36 | hCV1664153 | 49148151 | C | 0.47 | 0.80 | 0.74 | 0.87 | 0.82 |
| 37 | hCV9445425 | 49142230 | G | 0.43 | 0.51 | 0.38 | 0.35 | 0.49 |
| 38 | rs2899474 | 49141214 | T | 0.44 | 0.51 | 0.38 | 0.33 | 0.49 |
| 39 | rs749292 | 49137869 | A | 0.47 | 0.51 | 0.37 | 0.34 | 0.49 |
| 40 | hCV1203837 | 49136395 | C | 0.41 | 0.51 | 0.38 | 0.33 | 0.47 |
| 41 | hCV1138075 | 49130484 | G | 0.25 | 0.50 | 0.40 | 0.29 | 0.49 |
| 42 | hCV8234971 | 49128972 | A | 0.23 | 0.43 | 0.38 | 0.27 | 0.46 |
| 43 | rs1008805 | 49128737 | G | 0.15 | 0.43 | 0.40 | 0.60 | 0.38 |
| 44 | hCV8234947 | 49124592 | G | 0.15 | 0.43 | 0.39 | 0.28 | 0.49 |
| 45 | hCV8234935 | 49120798 | A | 0.13 | 0.45 | 0.39 | 0.31 | 0.52 |
| 46 | rs767199 | 49119525 | A | 0.13 | 0.44 | 0.37 | 0.31 | 0.52 |
| 47 | hCV11301451 | 49115160 | T | 0.15 | 0.45 | 0.36 | 0.32 | 0.50 |
| 48 | rs727479 | 49113685 | C | 0.16 | 0.42 | 0.32 | 0.42 | 0.31 |
| 49 | hCV8234874 | 49113193 | T | 0.17 | 0.42 | 0.32 | 0.42 | 0.33 |
| 50 | rs2414096 | 49108917 | A | 0.17 | 0.45 | 0.38 | 0.34 | 0.52 |
| 51 | rs700518 | 49108250 | C | 0.19 | 0.47 | 0.40 | 0.37 | 0.54 |
| 52 | hCV8234838 | 49104311 | T | 0.29 | 0.42 | 0.32 | 0.42 | 0.33 |
| 53 | rs1065778 | 49099344 | C | 0.17 | 0.44 | 0.38 | 0.34 | 0.53 |
| 54 | hCV8234804 | 49096238 | T | 0.29 | 0.43 | 0.32 | 0.41 | 0.33 |
| 55 | hCV8234792 | 49091802 | G | 0.24 | 0.52 | 0.40 | 0.38 | 0.56 |
| 56 | hCV8234791 | 49090006 | C | 0.23 | 0.52 | 0.40 | 0.36 | 0.56 |
| 57 | rs1143704 | 49089840 | A | 0.23 | 0.52 | 0.40 | 0.36 | 0.56 |
| 58 | rs230463 | 49087258 | C | 0.22 | 0.53 | 0.40 | 0.36 | 0.57 |
| 59 | rs700519 | 49087106 | A | 0.16 | 0.04 | 0.27 | 0.04 | 0.03 |
| 60 | int7_14A | 49087012 | A | 0.06 | 0.01 | 0.01 | 0.17 | 0.09 |
| 61 | hCV8234767 | 49085131 | C | 0.21 | 0.51 | 0.40 | 0.37 | 0.57 |
| 62 | hCV8234755 | 49083949 | C | 0.25 | 0.52 | 0.39 | 0.36 | 0.57 |
| 63 | rs10046 | 49082124 | A | 0.24 | 0.52 | 0.40 | 0.36 | 0.56 |
| 64 | rs4646 | 49081982 | A | 0.30 | 0.37 | 0.32 | 0.41 | 0.28 |
| 65 | rs2255192 | 49079973 | T | 0.37 | 0.10 | 0.28 | 0.22 | 0.16 |
| 66 | rs934632 | 49074968 | A | 0.22 | 0.37 | 0.31 | 0.24 | 0.20 |
| 67 | rs879046 | 49071401 | C | 0.44 | 0.99 | 0.00 | 0.99 | 0.00 |

**Table 7.** (*Continued*)

| SNP no. | SNP ID | Position[a] | Minor allele | Minor allele frequency | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | African-Americans | Hawaiians | Japanese | Latinas | Whites |
| 68 | rs2899469 | 49061060 | A | 0.43 | 0.51 | 0.44 | 0.55 | 0.61 |
| 69 | rs934635 | 49057915 | A | 0.09 | 0.08 | 0.01 | 0.21 | 0.18 |
| 70 | rs2414094 | 49057449 | A | 0.44 | 0.49 | 0.44 | 0.57 | 0.61 |
| 71 | rs2414093 | 49057423 | A | 0.10 | 0.07 | 0.01 | 0.22 | 0.17 |
| 72 | rs745258 | 49055208 | C | 0.47 | 0.79 | 0.82 | 0.49 | 0.57 |
| 73 | rs2414092 | 49054804 | T | 0.12 | 0.07 | 0.01 | 0.21 | 0.16 |
| 74 | rs1122044 | 49051383 | C | 0.28 | 0.20 | 0.18 | 0.48 | 0.41 |

[a]SNP position is based on the April 2003 freeze of chromosome 15 (contig NT_010194, http://genome.ucsc.edu).

### Haplotype reconstruction and htSNP selection

Haplotype frequency estimates were constructed from genotype data in the multiethnic panel (one ethnicity at a time) within blocks using the expectation–maximization (E-M) algorithm of Excoffier and Slatkin (43). The squared correlation ($R_h^2$) between the true haplotypes ($h$) and their estimates from this calculation were then estimated as described by Stram *et al.* (35). Briefly, for any given set of true haplotype frequencies, $P_h$, we can make a formal calculation (under Hardy–Weinberg equilibrium) of the squared correlation, $R_h^2$, between the estimate, $E\{\delta_h(H_i)|G_i\}$, and the true value, $\delta_h(H_i)$, of the number of copies of $h$ carried by a randomly sampled subject [i.e. $\delta_h(H_i) = 0$, 1 or 2]. Here $G_i$ is the genotype data for each subject, $i$, and $H_i$ is the true (but generally unknown) pair of haplotypes carried by that individual. The estimate is calculated as

$$E\{\delta_h(H_i) \mid G_i\} = \frac{\Sigma_{H\sim Gi}\delta_h(H)p_{h1}p_{h2}}{\Sigma_{H\sim Gi}p_{h1}p_{h2}}$$

where $\Sigma_{H\sim Gi}$ indicates a summation over the haplotype pairs, $H = \{h_1, h_2\}$, that are compatible with the observed genotype data, and $p_h$ is the frequency of haplotype $h$.

Under an assumption of Hardy–Weinberg equilibrium (HWE), the correlation may be most easily calculated as

$$R_h^2 = \frac{\mathrm{Var}[E\{\delta_h(H_i) \mid G_i\}]}{2p_h(1-p_h)}$$

where the variance of the expectation is computed by averaging $E\{\delta_h(H)|G\}$ and $E\{\delta_h(H)|G\}^2$ over all possible genotypes $G$, weighting by the probability of each genotype. This method explicitly recognizes that it is genotypes rather than haplotypes that are directly read, taking account of the resulting haplotype uncertainty. This uncertainty has not generally been accounted for in other haplotype SNP picking methods (44,45). $R_h^2$ is a sample size inflation factor—to achieve equivalent power as having perfectly tagged the haplotypes using $N$ samples requires approximately $N/R_h^2$ samples.

htSNPs for the case–control study were then chosen by finding the minimum set of SNPs (within a block) which would have $R_h^2 \geq 0.7$ for all haplotypes with an estimated frequency of $\geq$5%. The actual $R_h^2 s$ achieved for the haplotypes defined are generally higher and are given in the Results section. A computer program (tagSNPs) for the calculation of $R_h^2$ is available at D. Stram's website (www-rcf.usc.edu/~stram). A total of 25 htSNPs were selected to distinguish the common haplotypes (frequencies $\geq$5%) in blocks 1–4 estimated in each ethnic group of the multiethnic panel. We included as htSNPs two well-studied sequence variants, *Arg264Cys* in exon 7 (rs700519) and an SNP in the 3′-UTR of exon 10 (rs10046), before minimizing the number of htSNPs required to predict the common haplotypes. We expected and observed only minor differences in haplotype frequencies predicted solely by the htSNPs versus haplotype frequencies as defined by all of the SNPs in the block based on the high $R_h^2 s$ for determining the common haplotypes (Fig. 1B).

In addition, we calculated the multivariate squared correlation, $R_s^2$, between measured and unmeasured SNPs as an alternative statistic not focused on haplotype prediction, but rather on 'reconstruction' of the unmeasured SNP genotypes exploiting the multivariate correlation between SNPs in a region of high LD. This correlation is computed, as is $R_h^2$, based on the estimated haplotype frequencies under HWE. The average $R_s^2$ value for each block was $\geq$0.97, showing that our choice of htSNPs provides good prediction of unmeasured SNPs as well as an optimal prediction of haplotypes.

### Genotyping in the case–control study

Genotyping of htSNPs in the case–control study was performed by the 5′ nuclease Taqman allelic discrimination assay using the ABI7900 (Applied Biosystems, Foster City, CA, USA) in the MEC Genotyping Laboratory and by MALDI-TOF using the Sequenom platform at the Whitehead Institute/MIT Center for Genome Research. We also evaluated the independent effect of the rare *Trp39Arg* missense variant located in exon 2 (rs2236722). Laboratory personnel were blinded to case–control status and ~5% of samples were included as duplicates. The concordance for the blinded samples was >99%.

### Comparison of haplotype frequencies between breast cancer cases and controls

Haplotype frequencies among breast cancer cases and controls were estimated using the htSNPs selected to distinguish the common haplotypes ($\geq$5%) for each ethnic group in the multiethnic panel. Following the method of Zaykin *et al.* (46), for each individual and each haplotype, $h$, the haplotype dosage estimate (i.e. an estimate of the number of copies of haplotype $h$) was computed using that individual's genotype data and haplotype frequency estimates obtained from the

combined (cases + controls) data set. These individual estimates were merged with all other individual-specific data. All the variables were used in unconditional logistic regression analyses with the estimate of haplotype dosage treated as a surrogate variable for the true haplotype. Under the null hypothesis (of no haplotype-specific effects on risk) the usual score test from the logistic regression, when haplotype is added to the model, will correspond to the test described by Zaykin *et al.* (46). We have found that this approach gives accurate estimates of the statistical significance (*P*-values), and that confidence intervals (CIs) are appropriate when $R^2_h$ is high (47). Odds ratios (ORs) and 95% CIs for each haplotype were estimated using the most common haplotype observed among all ethnic groups combined within each block as the reference category. Results were similar when evaluating each haplotype separately (versus all other haplotypes, data not shown). Analyses were stratified by ethnicity and a summary OR was estimated controlling for age and ethnicity. Results were also similar when adjusting for the established breast cancer risk factors (26), family history of breast cancer, body mass index, parity, age at first birth, age at menarche, menopausal status, type of menopause, age at menopause, use of hormone replacement therapy and alcohol consumption (data not shown). A likelihood ratio test was performed to globally test for associations with the common haplotypes in each block. For blocks where this global test was significant, we also formally tested for ethnic differences in haplotype-associated risks by performing a likelihood ratio test following the inclusion of an interaction term between the risk haplotypes and ethnicity in the multivariate model. One case missing age at diagnosis was removed from all analyses. Sixty-eight cases and 98 controls had high genotype failure rates due to low DNA concentration and were removed from all genetic analyses. We used the Statistical Analysis System for all analyses (48).

## ACKNOWLEDGEMENTS

We thank Loreall Pooler, David Wong, and Johannah Butler for their laboratory assistance and Mathew Freedman, Kristine Monroe, Hank Huang, Peggy Wan, John Casagrande, Stuart Wugalter, Faye Nagamine, Stephen Schaffner, Mark Daly, Stacey Gabriel and Siby Sebastian for their technical support. We are also indebted to the participants of the Multiethnic Cohort Study for their participation and commitment. This work was supported by National Cancer Institute grants CA 63464 and CA 54281 and a General Motors Cancer Research Scholar's Grant awarded to C.A.H.

## REFERENCES

1. Henderson, B.E., Ross, R.K., Pike, M.C. and Casagrande, J.T. (1982) Endogenous hormones as a major factor in human cancer. *Cancer Res.*, **42**, 3232–3239.
2. Thomas, H.V., Reeves, G.K. and Key, T.J. (1997) Endogenous estrogen and postmenopausal breast cancer: a quantitative review. *Cancer Causes Control*, **8**, 922–928.
3. Hankinson, S.E., Willett, W.C., Manson, J.E., Colditz, G.A., Hunter, D.J., Spiegelman, D., Barbieri, R.L. and Speizer, F.E. (1998) Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women. *J. Natl Cancer Inst.*, **90**, 1292–1299.
4. Means, G.D., Mahendroo, M.S., Corbin, C.J., Mathis, J.M., Powell, F.E., Mendelson, C.R. and Simpson, E.R. (1989) Structural analysis of the gene encoding human aromatase cytochrome P- 450, the enzyme responsible for estrogen biosynthesis. *J. Biol. Chem.*, **264**, 19385–19391.
5. Mahendroo, M.S., Means, G.D., Mendelson, C.R. and Simpson, E.R. (1991) Tissue-specific expression of human P-450AROM. The promoter responsible for expression in adipose tissue is different from that utilized in placenta. *J. Biol. Chem.*, **266**, 11276–11281.
6. Means, G.D., Kilgore, M.W., Mahendroo, M.S., Mendelson, C.R. and Simpson, E.R. (1991) Tissue-specific promoters regulate aromatase cytochrome P450 gene expression in human ovary and fetal tissues. *Mol. Endocrinol.*, **5**, 2005–2013.
7. Sasano, H., Nagura, H., Harada, N., Goukon, Y. and Kimura, M. (1994) Immunolocalization of aromatase and other steroidogenic enzymes in human breast disorders. *Hum. Pathol.*, **25**, 530–535.
8. Utsumi, T., Harada, N., Maruta, M. and Takagi, Y. (1996) Presence of alternatively spliced transcripts of aromatase gene in human breast cancer. *J. Clin. Endocrinol. Metab.*, **81**, 2344–2349.
9. Yue, W., Wang, J.P., Hamilton, C.J., Demers, L.M. and Santen, R.J. (1998) *In situ* aromatization enhances breast tumor estradiol levels and cellular proliferation. *Cancer Res.*, **58**, 927–932.
10. Maggiolini, M., Bonofiglio, D., Pezzi, V., Carpino, A., Marsico, S., Rago, V., Vivacqua, A., Picard, D. and Ando, S. (2002) Aromatase overexpression enhances the stimulatory effects of adrenal androgens on MCF7 breast cancer cells. *Mol. Cell Endocrinol.*, **193**, 13–18.
11. van Landeghem, A.A., Poortman, J., Nabuurs, M. and Thijssen, J.H. (1985) Endogenous concentration and subcellular distribution of estrogens in normal and malignant human breast tissue. *Cancer Res.*, **45**, 2900–2906.
12. Harada, N., Utsumi, T. and Takagi, Y. (1993) Tissue-specific expression of the human aromatase cytochrome P-450 gene by alternative use of multiple exons 1 and promoters, and switching of tissue-specific exons 1 in carcinogenesis. *Proc. Natl Acad. Sci. USA*, **90**, 11312–11316.
13. Agarwal, V.R., Bulun, S.E., Leitch, M., Rohrich, R. and Simpson, E.R. (1996) Use of alternative promoters to express the aromatase cytochrome P450 (*CYP19*) gene in breast adipose tissues of cancer-free and breast cancer patients. *J. Clin. Endocrinol. Metab.*, **81**, 3843–3849.
14. Goss, P.E. and Strasser, K. (2001) Aromatase inhibitors in the treatment and prevention of breast cancer. *J. Clin. Oncol.*, **19**, 881–894.
15. Bonneterre, J., Thurlimann, B., Robertson, J.F., Krzakowski, M., Mauriac, L., Koralewski, P., Vergote, I., Webster, A., Steinberg, M. and von Euler, M. (2000) Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: results of the Tamoxifen or Arimidex Randomized Group Efficacy and Tolerability study. *J. Clin. Oncol.*, **18**, 3748–3757.
16. Nabholtz, J.M., Buzdar, A., Pollak, M., Harwin, W., Burton, G., Mangalik, A., Steinberg, M., Webster, A. and von Euler, M. (2000) Anastrozole is superior to tamoxifen as first-line therapy for advanced breast cancer in postmenopausal women: results of a North American multicenter randomized trial. Arimidex Study Group. *J. Clin. Oncol.*, **18**, 3758–3767.
17. Mouridsen, H., Gershanovich, M., Sun, Y., Perez-Carrion, R., Boni, C., Monnier, A., Apffelstaedt, J., Smith, R., Sleeboom, H.P., Janicke, F. *et al.* (2001) Superior efficacy of letrozole versus tamoxifen as first-line therapy for postmenopausal women with advanced breast cancer: results of a phase III study of the International Letrozole Breast Cancer Group. *J. Clin. Oncol.*, **19**, 2596–2606.
18. Kristensen, V.N., Andersen, T.I., Lindblom, A., Erikstein, B., Magnus, P. and Borresen-Dale, A.L. (1998) A rare *CYP19* (aromatase) variant may increase the risk of breast cancer. *Pharmacogenetics*, **8**, 43–48.
19. Siegelmann-Danieli, N. and Buetow, K.H. (1999) Constitutional genetic variation at the human aromatase gene (*Cyp19*) and breast cancer risk. *Br. J. Cancer*, **79**, 456–463.
20. Probst-Hensch, N.M., Ingles, S.A., Diep, A.T., Haile, R.W., Stanczyk, F.Z., Kolonel, L.N. and Henderson, B.E. (1999) Aromatase and breast cancer susceptibility. *Endocr. Relat. Cancer*, **6**, 165–173.
21. Haiman, C.A., Hankinson, S.E., Spiegelman, D., De Vivo, I., Colditz, G.A., Willett, W.C., Speizer, F.E. and Hunter, D.J. (2000) A tetranucleotide repeat polymorphism in *CYP19* and breast cancer risk. *Int. J. Cancer*, **87**, 204–210.

22. Healey, C.S., Dunning, A.M., Durocher, F., Teare, D., Pharoah, P.D., Luben, R.N., Easton, D.F. and Ponder, B.A. (2000) Polymorphisms in the human aromatase cytochrome P450 gene (*CYP19*) and breast cancer risk. *Carcinogenesis*, **21**, 189–193.
23. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
24. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
25. Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
26. Pike, M.C., Kolonel, L.N., Henderson, B.E., Wilkens, L.R., Hankin, J.H., Feigelson, H.S., Wan, P.C., Stram, D.O. and Nomura, A.M. (2002) Breast cancer in a multiethnic cohort in Hawaii and Los Angeles: risk factor-adjusted incidence in Japanese equals and in Hawaiians exceeds that in whites. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 795–800.
27. Harada, N. (1988) Cloning of a complete cDNA encoding human aromatase: immunochemical identification and sequence analysis. *Biochem. Biophys. Res. Commun.*, **156**, 725–732.
28. Miyoshi, Y., Iwao, K., Ikeda, N., Egawa, C. and Noguchi, S. (2000) Breast cancer risk associated with polymorphism in *CYP19* in Japanese women. *Int. J. Cancer*, **89**, 325–328.
29. Nativelle-Serpentini, C., Lambard, S., Seralini, G.E. and Sourdaine, P. (2002) Aromatase and breast cancer: W39R, an inactive protein. *Eur. J. Endocrinol*, **146**, 583–589.
30. Kristensen, V.N., Anderson, T.I., Lindblom, L., Erikstein, B., Magnus, P. and Börrensen-Dale, A.L. (1998) A rare *CYP19* (aromatase) variant may increase the risk of breast cancer. *Pharmacogenetics*, **8**, 43–48.
31. Haiman, C.A., Hankinson, S.E., Spiegelman, D., Brown, M. and Hunter, D.J. (2002) No association between a single nucleotide polymorphism in *CYP19* and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 215–216.
32. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
33. Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L. and Nickerson, D.A. (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.*, **33**, 518–521.
34. Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
35. Stram, D.O., Haiman, C.A., Hirschhorn, J., Altshuler, D., Kolonel, L.N., Henderson, B.E., Pike, M.C. (2003) Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.*, **100**, 27–36.
36. Mullis, P.E., Yoshimura, N., Kuhlmann, B., Lippuner, K., Jaeger, P. and Harada, H. (1997) Aromatase deficiency in a female who is compound heterozygote for two new point mutations in the P450arom gene: impact of estrogens on hypergonadotropic hypogonadism, multicystic ovaries, and bone densitometry in childhood. *J. Clin. Endocrinol. Metab.*, **82**, 1739–1745.
37. Henderson, B.E. and Feigelson, H.S. (2000) Hormonal carcinogenesis. *Carcinogenesis*, **21**, 427–433.
38. Watanabe, J., Harada, N., Suemasu, K., Higashi, Y., Gotoh, O. and Kawajiri, K. (1997) Arginine-cysteine polymorphism at codon 264 of the human *CYP19* gene does not affect aromatase activity. *Pharmacogenetics*, **7**, 419–424.
39. Lee, K.M., Abel, J., Ko, Y., Harth, V., Park, W.Y., Seo, J.S., Yoo, K.Y., Choi, J.Y., Shin, A., Ahn, S.H. *et al.* (2003) Genetic polymorphisms of cytochrome P450 19 and 1B1, alcohol use, and breast cancer risk in Korean women. *Br. J. Cancer*, **88**, 675–678.
40. Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. and Kong, A. (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat. Genet.*, **21**, 213–215.
41. Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M., Wilkens, L.R., Pike, M.C., Stram, D.O., Monroe, K.R., Earle, M.E. and Nagamine, F.S. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.*, **151**, 346–357.
42. Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**, 49–67.
43. Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
44. Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**,7335–7339.
45. Ke, X. and Cardon, L.R. (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, **19**, 287–288.
46. Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J. and Ehm, M.G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, **53**, 79–91.
47. Stram, D.O., Pearce, C.L., Bretsky, P., Freedman, M., Hirschhorn, J., Altshuler, D., Kolonel, L.N. and Henderson, B.E. (2003) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum. Hered.* (in press).
48. SAS Institute (2001) *SAS release 8.02.* SAS Inc., Cary, NC.