Haplotype Analysis of *CYP11A1* Identifies Promoter Variants Associated with Breast Cancer Risk

Brian L. Yaspan,[1] Joan P. Breyer,[2] Qiuyin Cai,[2] Qi Dai,[2] J. Bradford Elmore,[2] Isaac Amundson,[2] Kevin M. Bradley,[2] Xiao-Ou Shu,[2] Yu-Tang Gao,[5] William D. Dupont,[3] Wei Zheng,[2] and Jeffrey R. Smith[1,2,4]

Departments of [1]Cancer Biology, [2]Medicine, and [3]Biostatistics, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine; [4]Medical Research Service, VA Tennessee Valley Healthcare System, Nashville, Tennessee; and [5]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China

Abstract

The *CYP11A1* gene encodes the cholesterol side chain cleavage enzyme that catalyzes the initial and rate-limiting step of steroidogenesis. A large number of epidemiologic studies have implicated the duration and degree of endogenous estrogen exposure in the development of breast cancer in women. Here, we conduct a systematic investigation of the role of genetic variation of the *CYP11A1* gene in breast cancer risk in a study of 1193 breast cancer cases and 1310 matched controls from the Shanghai Breast Cancer Study. We characterize the genetic architecture of the *CYP11A1* gene in a Chinese study population. We then genotype tagging polymorphisms to capture common variation at the locus for tests of association. Variants designating a haplotype encompassing the gene promoter are significantly associated with both increased expression (P = 1.6e−6) and increased breast cancer risk: heterozygote age-adjusted odds ratio (OR), 1.51 [95% confidence interval (95% CI), 1.19–1.91]; homozygote age-adjusted OR, 2.94 (95% CI, 1.22–7.12), test for trend, P = 5.0e−5. Among genes controlling endogenous estrogen metabolism, *CYP11A1* harbors common variants that may influence expression to significantly modify risk of breast cancer. [Cancer Res 2007;67(12):5673–82]

Introduction

Endogenous estrogen exposure in women is strongly associated with risk of breast cancer (1, 2). Genetic variation within genes encoding enzymes of the biosynthetic pathway could greatly influence estrogen exposure and associated breast cancer risk (3). The conversion of cholesterol to pregnenolone is the common initial step in the biosynthesis of sex hormones, including estrogen, progesterone, and androgens. This rate-limiting conversion is catalyzed in steroidogenic tissues by the cholesterol side chain cleavage enzyme, the Cyp11A cytochrome P450 (4). We previously showed significant allelic association of a simple tandem repeat (STR) polymorphism upstream of the *CYP11A1* gene with breast cancer risk within a Chinese study population (5). Linkage and allelic association at the marker has also been observed in the androgen-related polycystic ovary syndrome (6). This STR is a pentanucleotide repeat (D15S520 at 15q24.1, $[TAAAA]_n$) located 487 bp upstream of the first exon of *CYP11A1*, a region not conserved between human and mouse. Three major alleles of four, six or eight repeats account for nearly all variation at the marker among Chinese. The eight-repeat allele is associated with a dose-dependant elevated risk of breast cancer [heterozygote odds ratio (OR) = 1.5, 95% confidence interval (95% CI) 1.2–1.9; homozygote OR = 2.9, 95% CI 1.3–6.7; trend test P < 0.0001; ref. 5].

In this study, we sought to comprehensively characterize common genetic variation at *CYP11A1* to assess patterns of linkage disequilibrium (LD) and to refine our understanding of the contribution of *CYP11A1* genetic variation to breast cancer risk. The initial discovery of the single-allele association at one STR of the *CYP11A1* gene led us to hypothesize that it marked an uncharacterized haplotype conferring breast cancer risk. Among alleles of variant sites identifying a cancer-associated haplotype, a subset that directly marks it are candidates that may be functional in the disease. Those altering transcript expression or processing or the encoded enzyme itself remain of great interest in further delineating the role of this gene in common breast cancer. We tested this hypothesis within the Shanghai Breast Cancer Study using haplotype-based analyses. Here, we show that the disease-associated haplotype is designated by multiple variants upstream of the coding region. We further observe that *CYP11A1* expression in a lymphoblastoid cell line homozygous for the disease-associated haplotype is 2-fold greater than the expression in lymphoblastoid cell lines harboring alternative haplotypes. We conclude that common *cis*-acting variants upstream of the coding region may affect transcriptional regulation to influence breast cancer risk.

Materials and Methods

Study population. The Shanghai Breast Cancer Study has been previously described (5, 7). Briefly, study subjects were recruited between August 1996 and March 1998. All subjects were permanent residents of urban Shanghai without a prior history of any cancer and were alive at the time of interview. The study included 1,459 incident breast cancer cases diagnosed at an age between 25 and 64 years during the study period (91.1% of eligible cases). Cancer diagnoses for all patients were reviewed and confirmed by a panel of clinicians, including two senior pathologists. Unaffected controls were randomly selected from the general population using the Shanghai Resident Registry, a population registry containing demographic information for all residents of urban Shanghai. Inclusion criteria for controls were identical to those for cases with the exception of a breast cancer diagnosis. Controls were frequency matched on age (5-year intervals) to the expected age distribution of the case subjects in a 1:1 ratio.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Jeffrey R. Smith, Department of Medicine, Vanderbilt University School of Medicine, 529 Light Hall, 2215 Garland Avenue, Nashville, TN 37232-0275. Phone: 615-936-2171; Fax: 615-936-2296; E-mail: jeffrey.smith@vanderbilt.edu and Wei Zheng, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN. Phone: 615-936-0682; Fax: 615-322-1754; E-mail: wei.zheng@vanderbilt.edu.

doi:10.1158/0008-5472.CAN-07-0467

The study included 1,556 control subjects (90.3% of matched eligible controls). Blood samples for DNA extraction were collected from 1,193 (82%) cases and 1,310 (84%) controls. For these subjects, the mean age was 47.5 for cases and 47.1 for controls. Mean age of menarche was 14.5 for cases and 14.7 for controls. Mean age of menopause was 48.2 for cases and 47.5 for controls. Among cases, 68% were premenopausal. Breast cancer was observed in first-degree relatives of 3.4% of cases and 2.4% of controls. All study participants provided written informed consent under an approved institutional review board protocol.

To preserve the limited DNA from study subjects recruited in the Shanghai Breast Cancer Study, allele discovery used DNAs obtained from Chinese cell lines of the Coriell Institute for Medical Research (Camden, NJ). These included NA18524, NA18526, NA18529, NA18532, NA18537, NA18540, NA18542, NA18545, NA18547, NA18550, NA18552, NA18555, NA18558, NA18561, NA18562, NA18563, NA18564, NA18566, NA18570, NA18571, NA18572, NA18573, NA18576, NA18577, NA18579, NA18582, NA18592, NA18593, NA18594, NA18603, NA18605, NA18608, NA18609, NA18611, NA18612, NA18620, NA18621, NA18622, NA18623, NA18624, NA18632, NA18633, NA18636, NA18637, NA00576, NA03433, NA13411, NA14821, NA16654, NA16688, NA16689, NA17013, NA17014, NA17015, NA17016, NA17017, NA17018, NA17019, and NA17020.

Variant discovery and confirmation. To capture genetic diversity of *CYP11A1*, database single-nucleotide polymorphisms (SNP) annotated in dbSNP were screened for common polymorphism in the study population. Fifty-three annotated SNPs spanning *CYP11A1* from 7.8 kb 5′ of the 29.9-kb gene to 10 kb 3′ were genotyped to assess polymorphism. This was done in quadruplicate among 15 Chinese cell line DNAs. The screening set was estimated to provide 95% power to detect a polymorphism with a minor variant frequency of 0.10 and 78% power with a frequency of 0.05.

The 15 Chinese cell lines were also used for *de novo* SNP discovery by dual single-strand conformational polymorphism (SSCP) methods and resequencing. Where either SSCP method identified a variant, conformers were resequenced for allele discovery. Overlapping amplimers across *CYP11A1* were used for polymorphism screening. This included 142 amplimers spanning from 1.9 kb 5′ to 98 bp 3′ of the gene. Intron 1 of the *CYP11A1* gene contains a 13.8-kb nearly contiguous interval of RepBase repeats. Select nonunique regions embedded within that interval (presenting an impediment for PCR-based assay) were omitted from the survey, as outlined in Fig. 1.

Characterization of haplotypes and LD was conducted among a pilot subset of subjects of the Shanghai Breast Cancer Study that included 178 cases and 178 controls. Subsequent genotyping of tagging SNPs and STRs for tests of association with breast cancer was conducted among 1,159 cases and 1,236 controls.

To identify additional genetic variation on the disease-associated haplotype, Chinese cell line GM16654, which is homozygous for the disease haplotype, and comparative cell line GM10859 (CEPH 1347-02) were resequenced from 1.9 kb 5′ to 98 bp 3′ of the gene (again omitting nonunique regions within the 13.8-kb interval of intron 1) as outlined in Fig. 1. All exons and exon-intron junctions were additionally resequenced in five subjects of the Shanghai Breast Cancer Study, harboring five common haplotypes, including a subject homozygous for the disease-associated haplotype. Sequencing used BigDye terminator chemistry on a 3100 Genetic Analyzer (Applied Biosystems). Discovered variants were then genotyped in the 59 Chinese cell line DNAs to assign alleles to known haplotypes.

Fifteen novel SNPs discovered in the study have been submitted to dbSNP (ss68316999–ss68317010 and ss68362647–ss68362649). Two novel polymorphic STRs have been submitted to GDB and to dbSNP (D15S1547/ss69363921 and D15S1546/ss69363922).

SNP genotyping. We genotyped SNPs by single-nucleotide primer extension and fluorescence polarization in 384-well format (8). Reaction processing entailed three steps: a 4.4-μL PCR reaction, addition of 4 μL of an exonuclease I (New England Biolabs) and calf intestinal alkaline phosphatase (Promega) reagent mix to degrade unincorporated primer and dephosphorylate deoxynucleotide triphosphates (dNTP), and a final addition of 4 μL of an Acyclopol and Acycloterminator reagent mix for the primer extension reaction (AcycloPrime FP SNP Detection System, Perkin-Elmer). Each PCR mixture included 0.1 unit of AmpliTaq Gold DNA polymerase, 1× Buffer II (Applied Biosystems), 2.5 mmol/L $MgCl_2$, 0.25 mmol/L dNTP, 335 nmol/L of each primer, and 2 ng DNA template. We detected incorporation of R110-labeled and TAMRA-labeled terminators by fluorescence polarization on a Molecular Devices LJL Analyst HT. Both forward and reverse strand extension primers were tested to select the most robust assay. Amplimer and extension primer sequences for genotyped SNPs of Fig. 2 are provided in Supplementary Table S1.

STR genotyping. 5′-Dye-labeled fluorescent amplimers were detected on an ABI PRISM 3700 (Applied Biosystems). Primers were designed using a tailing strategy to promote full nontemplated nucleotide addition by AmpliTaq Gold DNA polymerase (Applied Biosystems), providing unambiguous detection of alleles separated by 1 bp (9). PCR conditions were as described above. Primer sequences for polymorphic STRs are provided in Supplementary Table S1. Allele fragment size estimation was accomplished using the internal size standard Genescan 400HD ROX and the local Southern algorithm of GENESCAN software. Editing of alleles was done in GENOTYPER (Applied Biosystems).

Single-stranded conformation polymorphism detection. Amplimers were electrophoresed on 0.5× mutation detection enhancement gels (Cambrex Biosciences) at room temperature at 2 W for 14 h and at 4°C at 4 W for 14 h. PCR conditions were as described above. Amplimers were visualized by silver staining (10). Representative conformers were sequenced using BigDye terminator chemistry on a 3100 Genetic Analyzer (Applied Biosystems) to identify the polymorphic sites.

Statistical analyses. Hardy-Weinberg equilibrium (HWE) for markers was calculated using the Stata package *genassoc* of Clayton (11). Pairwise LD between SNPs was calculated and visualized using Haploview version 3.2 (12). Pairwise LD for SNPs and multiallelic STRs were calculated and visualized using MIDAS version 1 (13). Tagging SNPs were selected using LDSelect with a minor allele frequency (MAF) threshold of 5% and an r^2 threshold of 0.7 (14). When multiple SNPs were assigned as tagging SNPs for a particular bin, the SNP with most robust assay performance was selected for that bin.

Population haplotype frequencies were estimated by the Bayesian method implemented in PHASE version 2.1 (15–17) and by an expectation-maximization (18) algorithm implemented in custom software that we based upon a parent program written by Fallin and Schork (19, 20). The custom expectation-maximization program enabled use of multiallelic markers, placed no hard-coded limit on the number of subjects or markers, and allowed parallel processing. Diplotypes were predicted using PHASE, and those predicted with a probability of >95% were used for tests of association.

The χ^2 test statistic was used to evaluate differences in allele or haplotype frequency of case and control groups. Alleles or haplotypes with an overall frequency of <0.05 were grouped for analysis. A sliding window approach tested a haplotype window of N markers, sliding the window along the map in single-marker increments (19, 21). For a given window of N adjacent markers, the profile of multiple common haplotypes and rare haplotypes as a group were evaluated in cases and controls by the χ^2 test statistic. Each N-marker haplotype and remaining haplotypes of the window as a group was also evaluated by the χ^2 test statistic. Permutation testing was used to assess significance. Subsequent estimation of effect size used logistic regression models adjusted for age (Intercooled Stata 9, Stata Corporation).

Cladistic modeling of haplotypes resolved by PHASE with ≥99% probability was accomplished using DNAPARS and DRAWTREE of the software package Phylip 3.6. The observed haplotype with the least number of state changes to all other observed haplotypes was designated as the outgroup for unrooted parsimony. Each multiallelic marker of N alleles was encoded as a series of N-1 binary allelic sites to allow inclusion in the model.

Expression analyses in lymphoblastoid cell lines. Expression analyses used RNA prepared from the lymphoblastoid cell lines GM16654, GM17020, and GM17014 (Coriell Institute for Medical Research) carrying select *CYP11A1* diplotypes. Cells were cultured at 37°C under 5% CO_2 in a medium containing RPMI 1640 with 2 mmol/L L-glutamine and 15% fetal

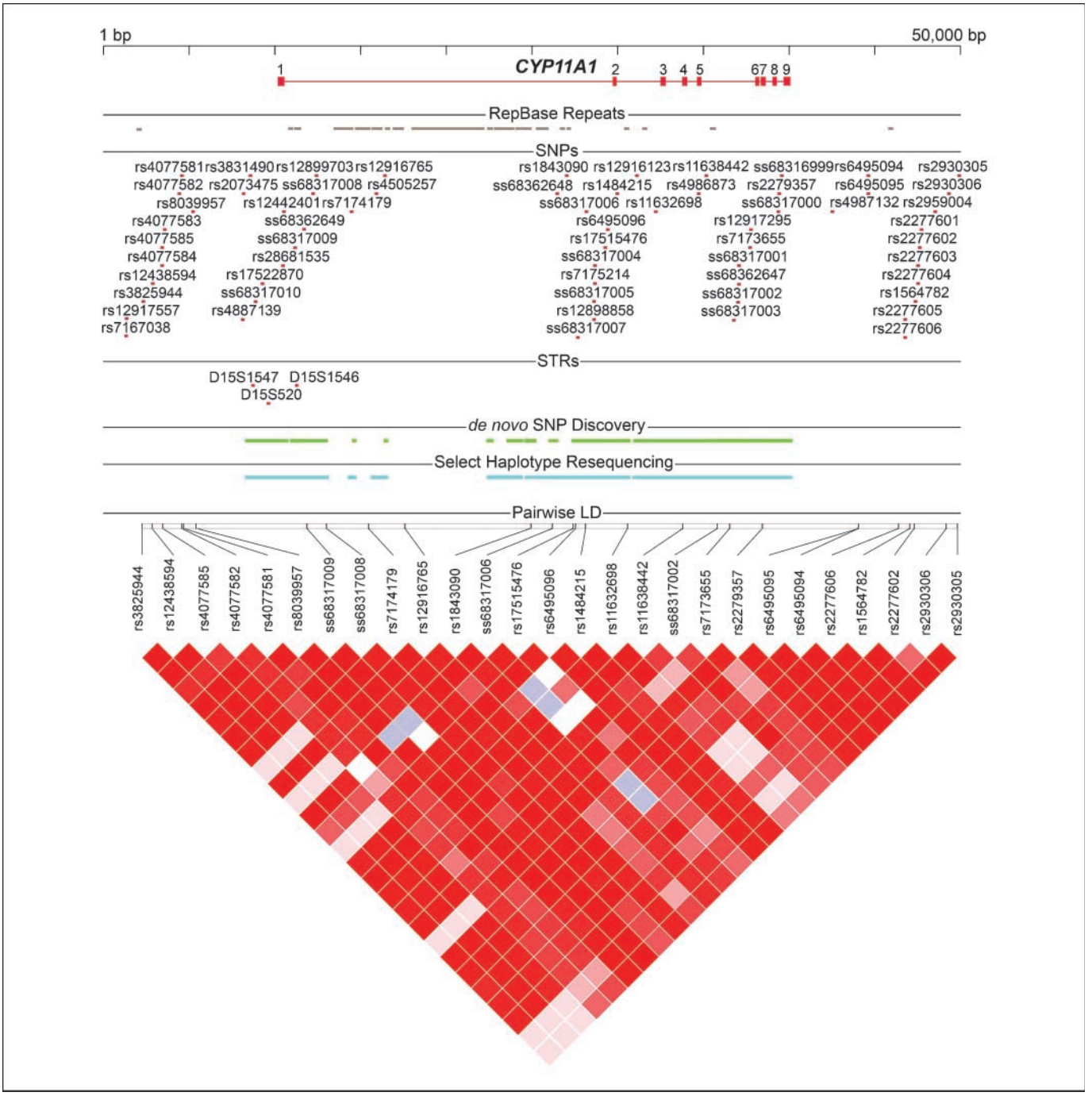
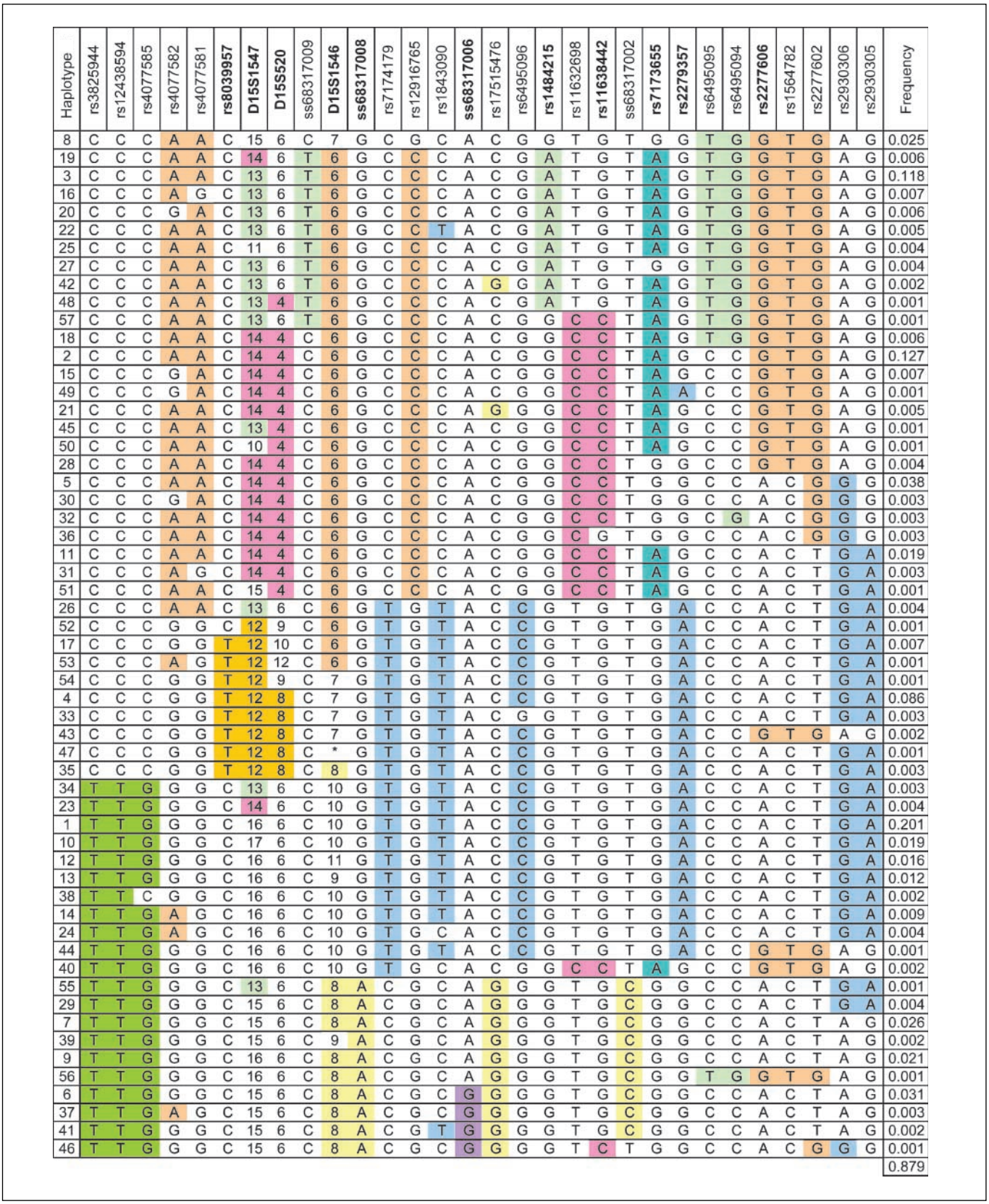


Figure 1. *CYP11A1* genetic architecture. A 50-kb interval from human chromosome 15q24.1 is depicted that encompasses the *CYP11A1* gene and 10 kb to each flank (National Center for Biotechnology Information build 36.1 from 72457199 to 72407199 bp). The gene's exons are numbered (*top*) with intervening introns. A 13.8-kb span of the first intron is dominated by repetitive elements. Variant sites observed among Chinese study subjects are positioned on the map. The 64 variants were identified by validation of sites annotated within dbSNP, by *de novo* discovery through SSCP/sequencing and by resequencing of select population haplotypes. A pairwise *D′* matrix (*bottom*) for 356 Chinese study subjects across a subset of 27 SNPs with MAF $\geq$ 0.05. The matrix graph indicates relatively strong LD across the locus. *Red, D′* = 1 (LOD $\geq$ 2); *blue, D′* = 1 (LOD < 2); *pink, D′* < 1 (LOD $\geq$ 2); *white, D′* < 1 (LOD < 2).

bovine serum. Total RNA from each cell line was prepared from cells in the log phase of growth using the RNeasy midi kit with on-column DNase treatment (Qiagen). RNA quality was assessed by reverse transcriptase PCR using two different sets of intron-spanning primers, one for phosphoglycerate kinase and one for p53, with a no-reverse transcriptase control to rule out DNA contamination. Nine 1-μg aliquots of total RNA of each cell line were reverse transcribed into single-stranded cDNA using High-Capacity cDNA Archive kit (Applied Biosystems). After cDNA synthesis, RNA was degraded by alkaline hydrolysis, pH was neutralized, cDNA was purified by adsorption to silica gel (QIAquick PCR Purification kit, Qiagen) and eluted in 60 μL of 10 mmol/L Tris Cl (pH 8.5). cDNA quantities were measured spectrophotometrically (NanoDrop ND-1000, NanoDrop Technologies).

A fluorescently labeled TaqMan MGB probe was used to quantify *CYP11A1* expression in each of the nine reverse-transcribed aliquots by real-time quantitative PCR. Each assay was done in quadruplicate. The probe spanned the exon 1 to exon 2 boundary within the coding region (Chr 15



Haplotype	rs3825944	rs12438594	rs4077585	rs4077582	rs4077581	**rs8039957**	**D15S1547**	**D15S520**	ss68317009	**D15S1546**	**ss68317008**	rs7174179	rs12916765	rs1843090	**ss68317006**	rs17515476	rs6495096	**rs1484215**	rs11632698	**rs11638442**	ss68317002	**rs7173655**	**rs2279357**	rs6495095	rs6495094	**rs2277606**	rs1564782	rs2277602	rs2930306	rs2930305	Frequency
8	C	C	C	A	A	C	15	6	C	7	G	C	G	C	A	C	G	G	T	G	T	G	G	T	G	G	T	G	A	G	0.025
19	C	C	C	A	A	C	14	6	T	6	G	C	C	C	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.006
3	C	C	C	A	A	C	13	6	T	6	G	C	C	C	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.118
16	C	C	C	A	G	C	13	6	T	6	G	C	C	C	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.007
20	C	C	C	G	A	C	13	6	T	6	G	C	C	C	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.006
22	C	C	C	A	A	C	13	6	T	6	G	C	C	T	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.005
25	C	C	C	A	A	C	11	6	T	6	G	C	C	C	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.004
27	C	C	C	A	A	C	13	6	T	6	G	C	C	C	A	C	G	A	T	G	T	G	G	T	G	G	T	G	A	G	0.004
42	C	C	C	A	A	C	13	6	T	6	G	C	C	C	A	G	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.002
48	C	C	C	A	A	C	13	4	T	6	G	C	C	C	A	C	G	A	T	G	T	A	G	T	G	G	T	G	A	G	0.001
57	C	C	C	A	A	C	13	6	T	6	G	C	C	C	A	C	G	G	C	C	T	A	G	T	G	G	T	G	A	G	0.001
18	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	T	G	G	T	G	A	G	0.006
2	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	G	T	G	A	G	0.127
15	C	C	C	G	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	G	T	G	A	G	0.007
49	C	C	C	G	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	A	C	C	G	T	G	A	G	0.001
21	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	G	G	G	C	C	T	A	G	C	C	G	T	G	A	G	0.005
45	C	C	C	A	A	C	13	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	G	T	G	A	G	0.001
50	C	C	C	A	A	C	10	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	G	T	G	A	G	0.001
28	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	G	G	C	C	G	T	G	A	G	0.004
5	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	G	G	C	C	A	C	G	G	G	0.038
30	C	C	C	G	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	G	G	C	C	A	C	G	G	G	0.003
32	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	G	G	C	G	A	C	G	G	G	0.003
36	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	G	T	G	G	C	C	A	C	G	G	G	0.003
11	C	C	C	A	A	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	A	C	T	G	A	0.019
31	C	C	C	A	G	C	14	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	A	C	T	G	A	0.003
51	C	C	C	A	A	C	15	4	C	6	G	C	C	C	A	C	G	G	C	C	T	A	G	C	C	A	C	T	G	A	0.001
26	C	C	C	A	A	C	13	6	C	6	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.004
52	C	C	C	G	G	C	12	9	C	6	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.001
17	C	C	C	G	G	T	12	10	C	6	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.007
53	C	C	C	A	G	T	12	12	C	6	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.001
54	C	C	C	G	G	T	12	9	C	7	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.001
4	C	C	C	G	G	T	12	8	C	7	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.086
33	C	C	C	G	G	T	12	8	C	7	G	T	G	T	A	C	G	G	T	G	T	G	A	C	C	A	C	T	G	A	0.003
43	C	C	C	G	G	T	12	8	C	7	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	G	T	G	A	G	0.002
47	C	C	C	G	G	T	12	8	C	*	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.001
35	C	C	C	G	G	T	12	8	C	8	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.003
34	T	T	G	G	G	C	13	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.003
23	T	T	G	G	G	C	14	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.004
1	T	T	G	G	G	C	16	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.201
10	T	T	G	G	G	C	17	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.019
12	T	T	G	G	G	C	16	6	C	11	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.016
13	T	T	G	G	G	C	16	6	C	9	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.012
38	T	T	C	G	G	C	16	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.002
14	T	T	G	A	G	C	16	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.009
24	T	T	G	A	G	C	16	6	C	10	G	T	G	C	A	C	C	G	T	G	T	G	A	C	C	A	C	T	G	A	0.004
44	T	T	G	G	G	C	16	6	C	10	G	T	G	T	A	C	C	G	T	G	T	G	A	C	C	G	T	G	A	G	0.001
40	T	T	G	G	G	C	16	6	C	10	G	T	G	C	A	C	G	G	C	C	T	A	G	C	C	G	T	G	A	G	0.002
55	T	T	G	G	G	C	13	6	C	8	A	C	G	C	A	G	G	G	T	G	C	G	G	C	C	A	C	T	G	A	0.001
29	T	T	G	G	G	C	15	6	C	8	A	C	G	C	A	G	G	G	T	G	C	G	G	C	C	A	C	T	G	A	0.004
7	T	T	G	G	G	C	15	6	C	8	A	C	G	C	A	G	G	G	T	G	C	G	G	C	C	A	C	T	A	G	0.026
39	T	T	G	G	G	C	15	6	C	9	A	C	G	C	A	G	G	G	T	G	C	G	G	C	C	A	C	T	A	G	0.002
9	T	T	G	G	G	C	16	6	C	8	A	C	G	C	A	G	G	G	T	G	C	G	G	C	C	A	C	T	A	G	0.021
56	T	T	G	G	G	C	16	6	C	8	A	C	G	C	A	G	G	G	T	G	C	G	G	T	G	G	T	G	A	G	0.001
6	T	T	G	G	G	C	15	6	C	8	A	C	G	C	G	G	G	G	T	G	C	G	G	C	C	A	C	T	A	G	0.031
37	T	T	G	A	G	C	15	6	C	8	A	C	G	C	G	G	G	G	T	G	C	G	G	C	C	A	C	T	A	G	0.003
41	T	T	G	G	G	C	15	6	C	8	A	C	G	T	G	G	G	G	T	G	C	G	G	C	C	A	C	T	A	G	0.002
46	T	T	G	G	G	C	15	6	C	8	A	C	G	C	G	G	G	G	T	C	T	G	G	C	C	A	C	G	G	G	0.001
																															0.879

Figure 2. *CYP11A1* haplotypes among 356 Chinese study subjects organized by cladistic similarity. Haplotypes (57) are predicted among subjects with a probability of ≥99%. Haplotypes are numbered in order of decreasing frequency. Each haplotype is designated by SNP allele and STR repeat count. *, an STR allele length other than a multiple of the repeat unit. Among the 27 SNPs (rs), those selected as tagging SNPs (*bold*), STRs (D15S#; *bold*). Alleles are color-coded to indicate membership in LD bins wherein pairwise r^2 values are ≥0.7.

72424438–72427449, assay Hs00167984_m1, Applied Biosystems). cDNA (5 ng) was amplified in a 5-μL reaction using the TaqMan system (Assays-On-Demand Gene Expression Products, TaqMan Universal PCR Master Mix, 7900HT Real-Time PCR System; Applied Biosystems). For each *CYP11A1* expression assay, results were normalized to the expression of the 18S rRNA housekeeping gene in the same sample (assay Hs99999901_s1, Applied Biosystems). Statistical comparisons were made using a one-way ANOVA and two-tailed Student's *t* test.

Results

We sought common polymorphisms at the *CYP11A1* gene by screening previously annotated variation and by *de novo* variant discovery within 30 chromosomes of Chinese cell lines. We tested SNPs annotated in dbSNP across an interval from 7.8 kb upstream to 10 kb downstream of the *CYP11A1* gene for polymorphism. We also sought previously undescribed common polymorphism through survey of the gene and ~2 kb 5′-flanking sequence by SSCP and resequencing. Repetitive sequence was an obstacle for unique assay. A 5.6-kb window of nonunique sequence of intron 1 and several additional small repetitive intronic regions totaling under 2 kb were omitted from SNP discovery efforts (Fig. 1). Collectively we identified 59 variant sites in the *CYP11A1* genomic region positioned on the map of Fig. 1. Of these, 80% were annotated in dbSNP. We developed assays for three STRs (including D15S1547, D15S520, and D15S1546, but omitting poly(A) tract indels rs3831490 and rs12899703) and 46 SNPs using Chinese cell lines. Among these markers, 3 STRs and 42 SNP assays were further genotyped in a subset of the Shanghai Breast Cancer Study population for the assessment of MAF, HWE, and haplotype diversity and for the selection of tagging markers. This study population subset included 178 cases and 178 controls. This yielded 3 polymorphic STRs and 27 SNPs (Fig. 2) with MAFs $\geq$ 0.05 and in HWE ($P \geq 0.05$) for inclusion in analyses. These SNPs had MAFs that ranged from 0.49 to 0.06 among controls. STR heterozygosities were 0.79 (D15S1547), 0.52 (D15S520), and 0.70 (D15S1546).

Pairwise LD across the *CYP11A1* gene was relatively strong in the study population and without clear LD block subdivision. A Haploview plot of SNP allele pairwise D' values is presented in Fig. 1. If an STR was highly mutable, one would anticipate low LD with neighboring SNPs. Instead, specific alleles of the STRs were in strong LD with select SNP alleles and efficiently tagged SNP haplotypes with few assays (Fig. 2). For example, the T allele of SNP rs8039957 (associated with breast cancer risk as shown further below) had pairwise D' values of 0.93 with the 12-repeat allele of D15S1547, 0.86 with the 8-repeat allele of D15S520, and 0.58 with the 8-repeat allele of D15S1546. Throughout the article, we refer to each SNP allele as that on the coding strand of the chromosome.

We sought an efficient set of tagging markers among the 3 STRs and 27 SNPs to capture *CYP11A1* gene diversity for tests of association with breast cancer. Eight SNPs and three STRs were selected as robust tagging markers, each with an allele in pairwise LD with an allele of remaining markers with an $r^2 \geq 0.70$ for the control group. This set of markers included rs8039957, D15S1547, D15S520, D15S1546, ss68317008, ss68317006, rs1484215, rs11638442, rs7173655, rs2279357, and rs2277606. Four SNPs at map ends (rs3825944, rs12438594, rs4077585, and rs2930306) were less efficiently tagged by the set, with maximal r^2 values ranging from 0.57 to 0.66.

Diplotypes of the 356 Shanghai Breast Cancer Study subjects were inferred for frequency estimation. Figure 2 illustrates haplotypes inferred by PHASE with a probability of $\geq$0.99; these are presented in an order predicted by cladistic modeling. Each haplotype has an identifying number from 1 to 57 (assigned by order of decreasing haplotype frequency). These account for 88% of all *CYP11A1* haplotypes in this population. Only five haplotypes were present with greater than a frequency of 0.05.

The STR alleles marked predominant SNP haplotypes well, in concordance with the high-measured pairwise LD values. Among more closely related SNP haplotypes (proximal in Fig. 2), STR alleles do deviate from the principal one and tend to do so by one- or two-repeat increments. This may reflect a stepwise rather than stochastic mutational mechanism (22, 23). Typical STR polymorphisms have alleles varying in increments of the repeat unit. D15S1547 is a dimer, D15S520 is a pentamer, and D15S1546 is a tetramer. However, D15S520 is distinct because predominant population alleles are in increments of 10 bp rather than 5 bp. We subcloned and resequenced each of the major alleles to confirm this.

We next genotyped the set of 11 tagging markers in 1159 breast cancer cases and 1236 controls of Shanghai Breast Cancer Study population to explore *CYP11A1* contribution to breast cancer risk. Data was obtained on 94% of genotypes sought (per marker range, 86–98%). Each of the tagging SNPs and STRs was in HWE ($P \geq 0.05$). Table 1 presents single-allele association results, comparing the case and control groups for these markers. The most significant evidence of association is observed at the three most 5′ markers, each just upstream of the *CYP11A1* coding region. Significance estimates by permutation testing range from $P = 2.0e{-}5$ to $4.1e{-}4$ for one allele at each of these markers. Each of the risk alleles observed in single-allele association tests (the T allele of rs8039957, 12-repeat allele of D15S1547, 8-repeat allele of D15S520, and 8-repeat allele of D15S1546) marks closely related haplotypes in the 5′ end of the *CYP11A1* gene, predominated in prevalence by haplotype 4 of Fig. 2 (frequency, 0.086).

We explored haplotype association using a sliding window approach across the tagging marker *CYP11A1* map. This implicates a haplotype over the 5′ region of the *CYP11A1* gene in breast cancer risk. Figure 3 presents haplotype association results for a window of three adjacent markers that is moved across the gene map in single-marker increments. The overall analysis was of windows ranging in size from 2 to 11 markers. Within each window, we inferred case and control haplotype frequencies by two independent methods: (*a*) by estimation of group frequencies using the expectation-maximization algorithm and (*b*) by assignment of individual study subject diplotype using PHASE. The two approaches yielded fully concordant results. Tests included those assessing overall haplotype profile differences between cases and controls (e.g., Fig. 3*A*), as well as those assessing excess of a given haplotype among cases relative to controls (e.g., Fig. 3*B*). A significant overall haplotype frequency profile difference between case and control groups was observed for all windows of width at two to six markers that included any of the three most 5′ markers. Significant overall haplotype profile differences were observed for 70% of windows of any width that included at least one of these markers (peak $P = 4.2e{-}4$). A total of 55 windows were evaluated. These multiple comparisons were not independent; thus, the Bonferroni corrected $P = 0.023$ is conservative. Within each significant window, individual haplotype comparisons were uniformly consistent with an excess of haplotype 4 (Fig. 2) in cases relative to controls (peak $P = 1.6e{-}5$, conservatively corrected by the factor of 585 haplotypes tested at the 55 windows to $P = 0.009$). These analyses identify the promoter region of the *CYP11A1* gene as a source of breast cancer risk in the study population.

Table 1. *CYP11A1* alleles and breast cancer risk

Marker	Allele	Cases, *n* (%)	Controls, *n* (%)	*P*
rs8039957	C	1776 (86.1)	2014 (90.0)	5.9e−5
	T	288 (14.0)	224 (10.0)	
D15S1547	12	292 (13.8)	226 (10.3)	4.1e−4
	13	356 (16.8)	415 (19.0)	0.063
	14	501 (23.7)	530 (24.2)	0.642
	15	267 (12.6)	281 (12.8)	0.783
	16	644 (30.4)	676 (30.9)	0.716
	Others	58 (2.7)	60 (2.7)	
		Overall χ^2 = 13.8 (P = 0.017)		
D15S520	4	528 (23.4)	550 (23.1)	0.807
	6	1410 (62.4)	1579 (66.3)	0.005
	8	292 (12.9)	219 (9.2)	2.0e−5
	Others	28 (1.2)	32 (1.4)	
		Overall χ^2 = 17.5 (P = 6.1e−4)		
D15S1546	7	925 (42.5)	1012 (44.5)	0.164
	8	329 (15.1)	284 (12.5)	0.011
	9	281 (12.9)	276 (12.1)	0.414
	11	562 (25.8)	618 (27.2)	0.292
	Others	81 (3.6)	84 (3.7)	
		Overall χ^2 = 7.9 (P = 0.095)		
ss68317008	A	285 (12.8)	283 (11.9)	0.370
	G	1943 (87.2)	2087 (88.1)	
ss68317006	A	2117 (94.3)	2249 (95.6)	0.042
	G	127 (5.7)	103 (4.4)	
rs1484215	A	380 (17.2)	455 (19.5)	0.042
	G	1836 (82.9)	1883 (80.5)	
rs11638442	C	569 (26.0)	573 (24.8)	0.355
	G	1623 (74.0)	1741 (75.2)	
rs7173655	A	797 (35.9)	878 (37.9)	0.166
	G	1421 (64.1)	1438 (62.1)	
rs2279357	A	947 (42.7)	963 (41.6)	0.416
	G	1269 (57.3)	1353 (58.4)	
rs2277606	A	1347 (62.3)	1342 (59.2)	0.031
	G	815 (37.7)	926 (40.8)	

We used logistic regression adjusted for age to assess the effect size of haplotype 4 relative to other haplotypes as a group. We evaluated the upstream promoter region of haplotype 4, delineated by markers rs8039957, D15S1547, and D15S520. The resulting estimates for the risk haplotype [T_12-repeat_8-repeat] are presented in Table 2. Inheritance of a single copy of the haplotype confers a 1.51-fold (95% CI, 1.19–1.91) significantly increased risk for breast cancer, and inheritance of two copies doubles this risk to 2.94-fold (95% CI, 1.22–7.12).

We reasoned that the list of potential functional candidate variants conferring disease risk would include (*a*) the alleles of the three markers above and (*b*) the alleles of other markers in strong LD with them, whether known or unknown. Among the 356 study subject subset, the maximum pairwise r^2 of alleles of any other known marker with the three alleles of interest was 0.23. In contrast, the allele T of rs8039957, 12-repeat of D15S1547, and 8-repeat of D15S520 had pairwise r^2 values ranging from 0.86 to 0.93, and each had weaker LD (r^2 range, 0.58–0.66) with the eight-repeat allele of D15S1546. Based upon direct sequencing data of the promoter region, one of the database-screened SNPs (rs4887139) that had failed assay development for the 356 subjects also potentially marked the disease-associated haplotype with a C allele. Additional unknown markers that might show strong LD with alleles of the disease-associated haplotype were of concern because the SSCP methods that we used for variant discovery at the *CYP11A1* gene lack complete sensitivity. Thus, we further searched for undiscovered variants that might also be functional candidates by resequencing the *CYP11A1* genomic region of a Chinese cell line and a Shanghai Breast Cancer Study case that we had characterized as homozygous for haplotype 4. These resequencing efforts identified five SNPs that had not previously been detected by SSCP or described in databases, two of which were not polymorphic in additional Chinese cell lines tested. The minor allele of a new polymorphic SNP in the first intron (now designated as rs12442401) seemed to directly mark the disease-associated haplotype. Data to support assignment of the minor allele to the disease haplotype was limited to that derived from sequencing, as we failed to develop a reliable genotyping assay for the marker. To summarize, the original three disease haplotype–marking variants and the additional rs4887139 and rs12442401 are candidates that may be functional in the phenotype. The disease-associated haplotype as defined by the full complement of observed variant sites is provided in Supplementary Table S2.

The evidence that these experiments uncovered supports a role for common *CYP11A1* promoter variation in breast cancer risk. Although *CYP11A1* expression is greatest in steroidogenic tissues, it

is also expressed in lymphocytes (24). Because we had identified the *CYP11A1* diplotype for each of 59 Chinese lymphoblastoid transformed cell lines, we subsequently evaluated expression of a cell line homozygous for the disease-associated haplotype (4 of Fig. 2) and compared with expression of two cell lines homozygous for alternative common haplotypes (1 and 3). *CYP11A1* expression was measured in total RNA prepared from the cell lines using a 5′ fluorogenic nuclease quantitative real-time PCR assay, normalizing to expression of 18S rRNA. Within these cell lines the expression of the disease-associated haplotype was roughly 2-fold greater than that of either alternative haplotype tested (Fig. 4). Increased relative expression is consistent with increased risk for breast cancer conferred by the promoter haplotype.

Discussion

We have conducted a detailed LD study of the *CYP11A1* gene and shown that a common promoter haplotype is associated with both

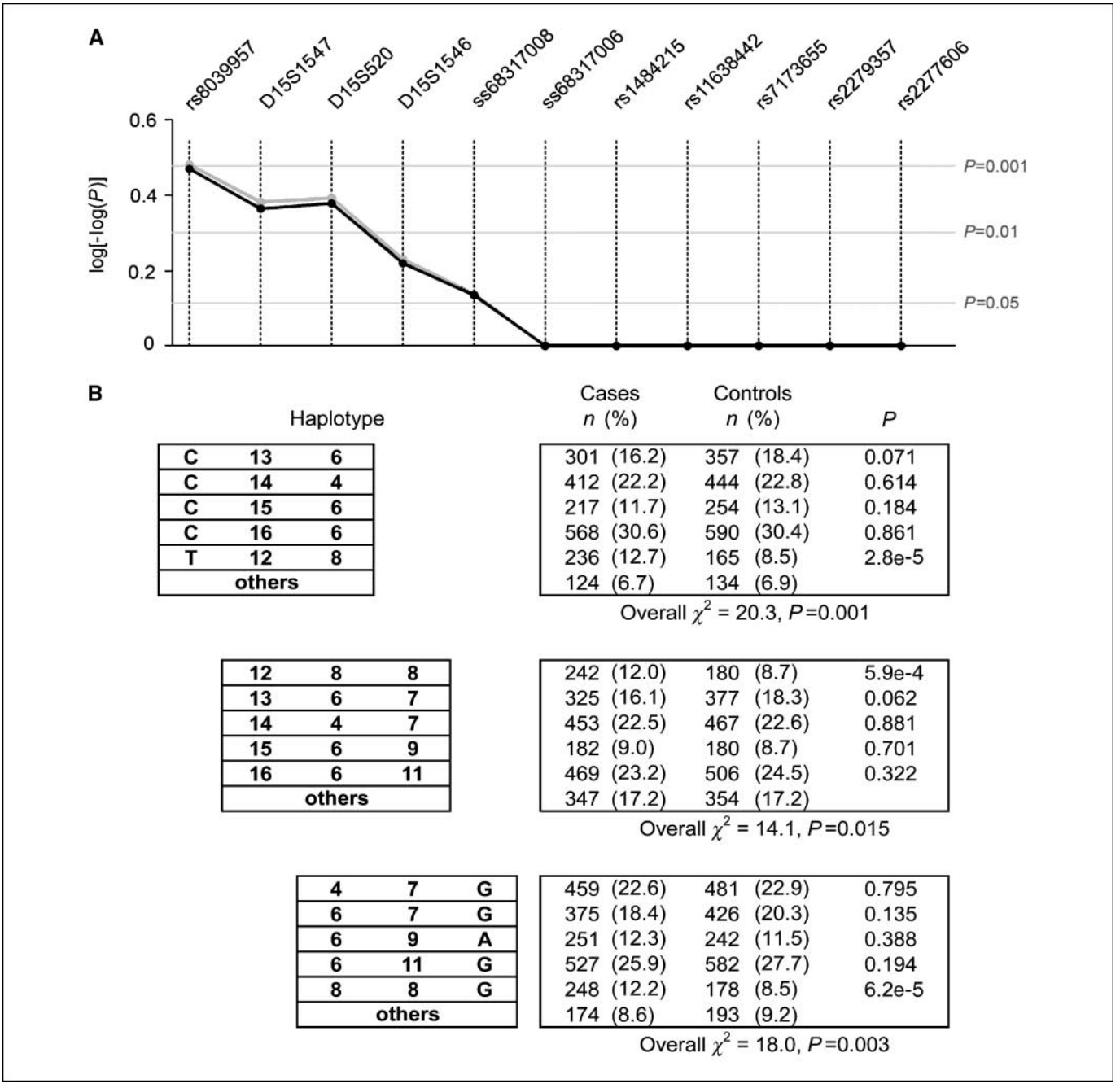


Figure 3. *CYP11A1* haplotypes and breast cancer risk. *A,* significance of the *CYP11A1* haplotype profile association with breast cancer for a window of three adjacent tagging markers, sliding across the map in single-marker increments. *Gray line,* case and control group frequency comparisons with haplotype-phase estimation by the expectation-maximization algorithm; *black line,* case and control group comparison in which individual study subject diplotypes were estimated by PHASE. *Points,* average of the log [−log(*P*)] transformed significance levels from overall χ^2 tests that included the marker. For reference, the transformed significance levels of $P = 0.05$, 0.01, and 0.001 are provided. *B,* evidence of association of each individual three-marker haplotype of the *CYP11A1* promoter region with breast cancer. Each haplotype is designated by tagging SNP allele and STR repeat count.

Table 2. *CYP11A1* promoter haplotype effect size upon breast cancer risk

Presence of haplotype T-12-8*	Cases, *n* (%)	Controls, *n* (%)	OR (95% CI) †	*P*
None	710 (88.5)	811 (91.6)	1.0 (reference)	
One copy	200 (10.5)	151 (8.0)	1.51 (1.19–1.91)	0.001
Two copies	18 (1.0)	7 (0.4)	2.94 (1.22–7.12)	0.017
Trend test				5.0e−5

*Haplotype alleles of markers rs8039957, D15S1547, D15S520.
†OR adjusted for age.

increased expression and increased risk of breast cancer. Select alleles of three markers upstream of the coding region (rs8039957, D15S1547, D15S520) define the haplotype. Alleles of two additional nearby markers, rs4887139 and rs12442401, also potentially mark the haplotype of interest. An allele of D15S1546 of the first intron showed less LD with alleles of the associated haplotype and weaker association with breast cancer risk. As currently delineated, the etiologic haplotype resides in a small 4-kb to 5-kb region spanning the *CYP11A1* promoter and would have been detected in a HapMap-based study design by virtue of selection of tagging SNP rs8039957. In HapMap data of Chinese from Beijing, this SNP is in full LD with rs4887139 and with rs4278698 (a SNP that failed our assay design process).

Our observations are consistent with the important role of the cholesterol side-chain cleavage enzyme in steroid sex hormone biosynthesis and with epidemiologic studies implicating estrogen biosynthesis and metabolism in breast cancer etiology (3). We estimate that population-attributable risk of the 5′ regulatory region haplotype of *CYP11A1* is 6.9%. This reflects an important contribution to breast cancer in the Chinese population. HapMap data for the CEU study population suggest a higher frequency of this haplotype (defined by rs8039957 allele T, rs4887139 allele C, and rs4278698 allele A) among Caucasians than among Chinese (25).

Because tissue-specific regulatory elements of *CYP11A1* are known to function in ovary and adrenal, it is conceivable that promoter haplotypes may be correlated selectively with premenopausal or postmenopausal breast cancer, reflecting the relative tissue origin of steroidogenesis. The cases of the Shanghai Breast Cancer Study that we evaluated are predominantly (68%) premenopausal, and evidence of association is strongest in this group (5). Intriguingly, Setiawan et al. also found evidence of association between a haplotype over the 5′ region of the *CYP11A1* gene and breast cancer risk in the Multiethnic Cohort Study (26). However, the risk haplotype that they identified (similar to haplotype 3 of Fig. 2) is distinct from the risk haplotype (4) identified in our study. Cases of the Multiethnic Cohort Study are predominantly (69%) postmenopausal. The risk haplotype of the Setiawan et al. study is tagged by rs3803463 at −7,542 bp upstream of the gene, a marker also in LD with rs1484215 between exons 2 and 3 (r^2 range, 0.75–0.87 in HapMap populations). In light of these collective findings, further epidemiologic evaluation of *CYP11A1* haplotypes in premenopausal and postmenopausal breast cancer, and investigation of their effect on tissue-specific expression is warranted.

The structure of the *CYP11A1* gene promoter has been extensively investigated in prior studies (27–39). The proximal promoter is composed of a TATA box, a highly conserved SF1/LRH-1 site, and two SP1 sites. Promoter deletion mapping has also identified a negative regulatory element residing between −300 and −660 bp (27, 33, 37). This region harbors nonconserved repetitive elements flanking D15S520 at −487 bp. The nonconserved CA simple-sequence repeat D15S1547 at −1,361 bp is also adjacent to a repetitive element. The upstream cyclic AMP (cAMP)–response sequence at −1,540 to −1,640 bp harbors two AP1/cAMP-responsive element binding protein binding sites flanking an SF1 site. Further upstream are two adrenal-specific enhancers between −1,840 and −1,900. SNPs marking the disease-associated haplotype, rs4887139 at −2,228, rs8039957 at −4,884 bp, and potentially rs4278698 at −4,984 do not reside within conserved regions. Both the $[AAAT]_n$ simple-sequence repeat of D15S1546 and SNP rs12442401 reside within nonconserved regions of the first intron. All identified variants of the disease-associated haplotype, thus, fall outside of conserved elements defined by vertebrate Multiz alignment (40) in the *CYP11A1* region. However, the

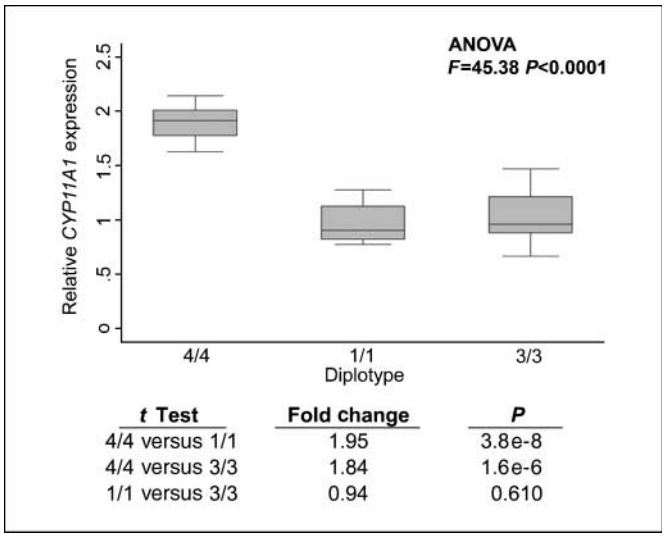


Figure 4. *CYP11A1* expression in lymphoblastoid cell lines. Expression within cell line GM16654 (homozygous for the breast cancer associated haplotype 4) is shown relative to that of cell lines GM17020 and GM17014 (homozygous for common haplotypes 1 and 3). *CYP11A1* expression is normalized to 18S rRNA levels. Each box plot presents nine independent measurements of expression within a given cell line (median, box range 25th–75th percentile, whiskers of data within 1.5-fold of the interquartile range). Significance is presented rejecting the hypothesis that all three expression levels are the same. Pairwise comparisons of cell line expression are also made, rejecting the hypothesis that expression of the 4/4 diplotype cell line is the same as that of each other cell line.

D15S520 repeat $[TAAAA]_n$ potentially resides within a described functional promoter element.

A $[TAAAA]_n$ polymorphic repeat has been shown to be a negative regulatory element within the promoter of the plasma sex hormone–binding globulin gene (*SHBG*; ref. 41); 6 to 11 repeats reside at −726 bp of that promoter. Reporter constructs carrying six repeats showed significantly less transcriptional activity than constructs carrying other repeat lengths. The six-repeat version of the *SHBG* promoter is also associated with lower SHBG levels (42). The six-repeat allele of D15S520 was the most commonly observed in our study, and two cell lines homozygous for the six-repeat allele each had significantly lower *CYP11A1* expression than a cell line homozygous for the disease-associated eight-repeat allele. The two promoter repeats may not be fully analogous, however, because increased risk of breast cancer in our study was associated only with the eight-repeat allele at *CYP11A1*, not with other non–six-repeat alleles. An even and odd number of repeats could alternatively orient closely flanking transcription factors on the same or opposite DNA helical faces to influence interactions; odd repeat alleles of D15S520 were relatively rare in both the Shanghai Breast Cancer Study and in the Multiethnic Cohort Study (26).

Heritable variation of both *cis* and *trans* regulatory elements controlling expression of steroid hormone biosynthesis and metabolism genes could greatly contribute to population breast cancer risk (43, 44). Broader investigation of this large network of genes should reasonably include genetic variation of potential regulatory elements. A genome-wide or a candidate gene association study based upon tagging SNP selection from current HapMap data could have detected association of *CYP11A1* with breast cancer risk in our study population. A direct investigation of *SHBG* promoter variation in breast cancer risk has not yet been conducted, although higher plasma levels of SHBG (with corresponding lower levels of circulating estrogen) have been associated with reduced risk for breast cancer (45). Among other genes of the steroid hormone regulatory network, a SNP within the human progesterone receptor gene promoter, located between its two alternative isoform transcript start sites, has been shown to have a direct effect on expression (46). That promoter variant was further associated with breast cancer risk in the Nurses' Health Study (46). Systematic investigation of steroid hormone biosynthesis and metabolism gene variation may provide a more comprehensive picture of the role of these pathways in breast cancer risk.

Acknowledgments

Received 2/5/2007; revised 4/6/2007; accepted 4/12/2007.

Grant support: U.S. Presidential Early Career Award for Scientists and Engineers (J.R. Smith). The Shanghai Breast Cancer Study was supported by National Cancer Institute grants R01 CA64277 and R01 CA90899 (W. Zheng).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank the study participants and staff of Shanghai Breast Cancer Study.

References

1. Travis RC, Key TJ. Oestrogen exposure and breast cancer risk. Breast Cancer Res 2003;5:239–47.
2. Folkerd EJ, Martin LA, Kendall A, Dowsett M. The relationship between factors affecting endogenous oestradiol levels in postmenopausal women and breast cancer. J Steroid Biochem Mol Biol 2006;102:250–5.
3. Mitrunen K, Hirvonen A. Molecular epidemiology of sporadic breast cancer. The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. Mutat Res 2003;544:9–41.
4. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–8.
5. Zheng W, Gao YT, Shu XO, et al. Population-based case-control study of 2CYP11A gene polymorphism and breast cancer risk. Cancer Epidemiol Biomarkers Prev 2004;13:709–14.
6. Gharani N, Waterworth DM, Batty S, et al. Association of the steroid synthesis gene CYP11a with polycystic ovary syndrome and hyperandrogenism. Hum Mol Genet 1997;6:397–402.
7. Gao YT, Shu XO, Dai Q, et al. Association of menstrual and reproductive factors with breast cancer risk: results from the Shanghai Breast Cancer Study. Int J Cancer 2000;87:295–300.
8. Chen X, Levine L, Kwok PY. Fluorescence polarization in homogeneous nucleic acid analysis. Genome Res 1999;9:492–8.
9. Brownstein MJ, Carpten JD, Smith JR. Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. Biotechniques 1996;20:1004–6, 8–10.
10. von Deimling A, Bender B, Louis DN, Wiestler OD. A rapid and non-radioactive PCR based assay for the detection of allelic loss in human gliomas. Neuropathol Appl Neurobiol 1993;19:524–9.
11. Cordell HJ, Clayton DG. Genetic association studies. Lancet 2005;366:1121–31.
12. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.
13. Gaunt TR, Rodriguez S, Zapata C, Day IN. MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. BMC Bioinformatics 2006;7:227.
14. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004; 74:106–20.
15. Marchini J, Cutler D, Patterson N, et al. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 2006;78:437–50.
16. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003;73:1162–9.
17. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–89.
18. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921–7.
19. Fallin D, Cohen A, Essioux L, et al. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 2001;11:143–51.
20. Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 2000;67:947–59.
21. Mathias RA, Gao P, Goldstein JL, et al. A graphical assessment of *P*-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. BMC Genet 2006;7:38.
22. Kayser M, Roewer L, Hedman M, et al. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 2000;66:1580–8.
23. Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 1998;62:1408–15.
24. Zhou Z, Shackleton CH, Pahwa S, White PC, Speiser PW. Prominent sex steroid metabolism in human lymphocytes. Mol Cell Endocrinol 1998;138:61–9.
25. The International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–320.
26. Setiawan VW, Cheng I, Stram DO, et al. A systematic assessment of common genetic variation in CYP11A and risk of breast cancer. Cancer Res 2006;66:12019–25.
27. Guo IC, Hu MC, Chung BC. Transcriptional regulation of CYP11A1. J Biomed Sci 2003;10:593–8.
28. Liu Z, Simpson ER. Steroidogenic factor 1 (SF-1) and SP1 are required for regulation of bovine CYP11A gene expression in bovine luteal cells and adrenal Y1 cells. Mol Endocrinol 1997;11:127–37.
29. Liu Z, Simpson ER. Molecular mechanism for cooperation between Sp1 and steroidogenic factor-1 (SF-1) to regulate bovine CYP11A gene expression. Mol Cell Endocrinol 1999;153:183–96.
30. Venepally P, Waterman MR. Two Sp1-binding sites mediate cAMP-induced transcription of the bovine CYP11A gene through the protein kinase A signaling pathway. J Biol Chem 1995;270:25402–10.
31. Ahlgren R, Suske G, Waterman MR, Lund J. Role of Sp1 in cAMP-dependent transcriptional regulation of the bovine CYP11A gene. J Biol Chem 1999;274:19422–8.
32. Doi J, Takemori H, Lin XZ, Horike N, Katoh Y, Okamoto M. Salt-inducible kinase represses cAMP-dependent protein kinase-mediated activation of human cholesterol side chain cleavage cytochrome P450 promoter through the CREB basic leucine zipper domain. J Biol Chem 2002;277:15629–37.
33. Chou SJ, Lai KN, Chung B. Characterization of the upstream sequence of the human CYP11A1 gene for cell type-specific expression. J Biol Chem 1996;271:22125–9.
34. Hu MC, Chiang EF, Tong SK, et al. Regulation of steroidogenesis in transgenic mice and zebrafish. Mol Cell Endocrinol 2001;171:9–14.
35. Huang Y, Hu M, Hsu N, Wang CL, Chung B. Action of hormone responsive sequence in 2.3 kb promoter of CYP11A1. Mol Cell Endocrinol 2001;175:205–10.

36. Chung BC, Guo IC, Chou SJ. Transcriptional regulation of the CYP11A1 and ferredoxin genes. Steroids 1997;62:37–42.
37. Guo IC, Tsai HM, Chung BC. Actions of two different cAMP-responsive sequences and an enhancer of the human CYP11A1 (P450scc) gene in adrenal Y1 and placental JEG-3 cells. J Biol Chem 1994;269:6362–9.
38. Guo IC, Chung BC. Cell-type specificity of human CYP11A1 TATA box. J Steroid Biochem Mol Biol 1999;69:329–34.
39. Monte D, DeWitte F, Hum DW. Regulation of the human P450scc gene by steroidogenic factor 1 is mediated by CBP/p300. J Biol Chem 1998;273:4585–91.
40. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–50.
41. Hogeveen KN, Talikka M, Hammond GL. Human sex hormone-binding globulin promoter activity is influenced by a (TAAAA)n repeat element within an Alu sequence. J Biol Chem 2001;276:36383–90.
42. Haiman CA, Riley SE, Freedman ML, Setiawan VW, Conti DV, Le Marchand L. Common genetic variation in the sex steroid hormone-binding globulin (SHBG) gene and circulating shbg levels among postmenopausal women: the Multiethnic Cohort. J Clin Endocrinol Metab 2005;90:2198–204.
43. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature 2005;437:1365–9.
44. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430:743–7.
45. Key T, Appleby P, Barnes I, Reeves G. Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. J Natl Cancer Inst 2002;94:606–16.
46. Huggins GS, Wong JY, Hankinson SE, De Vivo I. GATA5 activation of the progesterone receptor gene promoter in breast cancer cells is influenced by the +331G/A polymorphism. Cancer Res 2006;66:1384–90.