HHS Public Access

Author manuscript

Hum Genet. Author manuscript; available in PMC 2015 June 24.

Published in final edited form as:

Hum Genet. 2011 December ; 130(6): 767–775. doi:10.1007/s00439-011-1025-6.

A signature of balancing selection in the region upstream to the human *UGT2B4* gene and implications for breast cancer risk

Chang Sun[1], **Dezheng Huo**[2], **Catherine Southard**[1], **Barbara Nemesure**[3], **Anselm Hennis**[4], **Cristina Leske**[3], **Suh-Yuh Wu**[3], **David B. Witonsky**[1], **Olufunmilayo I. Olopade**[5], and **Anna Di Rienzo**[1]

[1]Department of Human Genetics, University of Chicago, 920 E. 58th Street, Chicago, IL 60637, USA

[2]Department of Health Studies, University of Chicago, Chicago, IL 60637, USA

[3]Department of Preventive Medicine, State University of New York at Stony Brook, Stony Brook, NY 11794, USA

[4]Faculty of Medical Sciences, University of the West Indies, Bridgetown, Barbados

[5]Department of Medicine, University of Chicago, Chicago, IL 60637, USA

Abstract

UDP-glucuronosyltransferase 2 family, polypeptide B4 (UGT2B4) is an important metabolizing enzyme involved in the clearance of many xenobiotics and endogenous substrates, especially steroid hormones and bile acids. The HapMap data show that numerous SNPs upstream of *UGT2B4* are in near-perfect linkage disequilibrium with each other and occur at intermediate frequency, indicating that this region might contain a target of natural selection. To investigate this possibility, we chose three regions (4.8 kb in total) for resequencing and observed a striking excess of intermediate-frequency alleles that define two major haplotypes separated by many mutation events and with little differentiation across populations, thus suggesting that the variation pattern upstream *UGT2B4* is highly unusual and may be the result of balancing selection. We propose that this pattern is due to the maintenance of a regulatory polymorphism involved in the fine tuning of *UGT2B4* expression so that heterozygous genotypes result in optimal enzyme levels. Considering the important role of steroid hormones in breast cancer susceptibility, we hypothesized that variation in this region could predispose to breast cancer. To test this hypothesis, we genotyped tag SNP rs13129471 in 1,261 patients and 825 normal women of African ancestry from three populations. The frequency comparison indicated that rs13129471 was significantly associated with breast cancer after adjusting for ethnicity [P = 0.003; heterozygous odds ratio (OR) 1.02, 95% confidence interval (CI) 0.81–1.28; homozygous OR 1.50, 95% CI 1.15–1.95]. Our results provide new insights into *UGT2B4* sequence variation and indicate that a signal of natural selection may lead to the identification of disease susceptibility variants.

Introduction

Glucuronidation is an important pathway for the clearance of numerous endobiotics and xenobiotics, including steroid hormones, bile acid, carcinogens and clinical drugs (Belanger

et al. 2003; Tukey and Strassburg 2000). This reaction can transfer the glucuronic acid from UDP glucuronic acid to available substrates and make them more water soluble and easier to be eliminated from the human body (Tukey and Strassburg 2000). In human, glucuronidation is performed by two main families of UDP-glucuronosyltransferases (UGTs), namely UGT1A and UGT2B (Guillemette 2003; King et al. 2000; Mackenzie et al. 2005). The latter one includes seven active members located in chromosome 4q13 (Guillemette 2003; King et al. 2000; Mackenzie et al. 2005) and among this family, the UGT2B4 gene is the most highly expressed member in the liver (Izukawa et al. 2009; Nakamura et al. 2008; Ohno and Nakajin 2009). In addition to liver, this gene is mainly expressed in heart, prostate, breast, and kidney (Gardner-Stephen and Mackenzie 2008; Nakamura et al. 2008; Ohno and Nakajin 2009). Despite the important biological role played by this gene, the impact of natural selection on patterns of variation has not been investigated.

Balancing selection has been invoked to explain patterns of variation in several human genes, including those coding for β-globin (Currat et al. 2002), glucose-6-phosphate dehydrogenase (*G6PD*; Verrelli et al. 2002), PTC (the bitter-taste receptor gene; Wooding et al. 2004), the sixth complement component (*C6*; Soejima et al. 2005), flavin-containing monooxygenase 3 (*FMO3*; Allerston et al. 2007), follicle stimulating hormone, beta polypeptide (*FSHB*; Grigorova et al. 2007), angiotensin-converting enzyme (*ACE*; Cagliani et al. 2010a) and, in particular, genes involved in immune response (Fumagalli et al. 2009b), such as endoplasmic reticulum aminopeptidase 1 and 2 (*ERAP1* and *ERAP2*; Andres et al. 2010; Cagliani et al. 2010b), leukocyte immunoglobulin-like receptor A3 (*LILRA3*; Hirayasu et al. 2006; Norman et al. 2004), chemokine C–C motif receptor 5 (*CCR5*; Bamshad et al. 2002), human leukocyte antigen (*HLA*; Black and Hedrick 1997; Liu et al. 2006; Tan et al. 2005), β-defensin (Cagliani et al. 2008; Hollox and Armour 2008), interleukin (IL)/IL receptor (Fumagalli et al. 2009c), and mediterranean fever (*MEFV*; Fumagalli et al. 2009a). Under balancing selection, alleles contributing to an adaptive phenotype are maintained at an equilibrium frequency as long as the selective pressure is acting (Andres et al. 2009; Charlesworth 2006). It has been proposed that it is a major mechanism for maintaining phenotypic and genetic variation in natural populations (Andres et al. 2009). If the selective pressures leading to a balanced polymorphism are steady for a long period of time, the expected signature in patterns of linked neutral variation is an excess of diversity relative to between-species divergence levels, a skew towards intermediate frequency alleles, and a distinct haplotype structure. This signature has been observed at several loci in the human genome, but it is not common (Bubb et al. 2006). This finding suggests that balancing selection in humans is rarely stable over long periods of time or that current genetic variation data sets are not well suited to reveal this signature.

Breast cancer is the most common malignant tumor in women and cumulative estrogen exposure over a life time has been proven to be a major risk factor (Cauley et al. 1999; Key 1999; Thomas et al. 1997). Due to the role of UGT2B enzymes in the metabolism of steroid hormones, it has been proposed that variation in the UGT2B genes, especially the ones that can reduce enzyme activity, is involved in breast cancer risk (Desai et al. 2003; Nagar and Remmel 2006). In *UGT2B4*, one SNP that causes an amino acid substitution (D458E) has

been tested for the effect on hormone levels and risk to breast cancer and was found not to be associated with these phenotypes (Sparks et al. 2004). Some non-coding SNPs which might regulate gene expression and enzyme abundance can also influence steroid hormone levels, thus constituting good candidates for breast cancer risk. However, no studies have been performed in *UGT2B4* non-coding variations so far.

Our preliminary bioinformatics analysis identified a large region upstream of the *UGT2B4* gene with a striking two-haplotype pattern and many SNPs at intermediate allele frequency (The International HapMap Consortium 2007; Fig. S1), raising the possibility that this region was a target of balancing selection. However, SNP ascertainment prevents an unbiased assessment of patterns of variation and an evaluation of the signature of balancing selection. Here, we conducted a re-sequencing survey of three segments in *UGT2B4* upstream region and identified high diversity levels, a strong excess of intermediate frequency variants, and two major haplotypes separated by a deep bifurcation, in all populations, consistent with balancing selection. In addition, we tested a SNP tagging the two major haplotypes in breast cancer case and control groups and found it to be significantly associated with disease risk.

Materials and methods

Samples and resequencing

Three segments in *UGT2B4* upstream region were randomly chosen for resequencing (Fig. S2). Fifty-six unrelated HapMap samples (24 YRI, 22 CEU and 10 ASN) were included in our resequencing survey. Amplification was performed using the primers in Table S1. After exonuclease I and Shrimp Alkaline Phosphatase (United States Biochemicals, Cleveland, OH) treatment, sequencing was performed using internal primers in Table S1 and BigDye Terminator v3.1 (Applied Biosystems, Foster City, CA, USA). In total, 4.8 kb were amplified and resequenced. Polymorphisms were scored by PolyPhred (Stephens et al. 2006) and inspected visually. Visual genotype and LD plots were drawn using the Genome Variation Server (http:// gvs. gs. washington. edu/ GVS/). In addition to the human samples, the homologous counterpart of segments 2 and 3 was also amplified and resequenced in ten unrelated chimpanzees (*Pan troglodytes*; from the Yerkes National Primate Research Center, http:// www. yerkes. emory. edu/ index. html) using the primers in Table S2.

Data analysis

Population genetics indices, including segregating sites (S), nucleotide diversity (π; Tajima 1983), Watterson's estimator of the population mutation rate parameter (θ_w; Watterson 1975), Tajima's D (Tajima 1989), and F_{ST} (Weir and Cockerham 1984; Wright 1950) were determined by Slider (http:// genapps. uchicago. edu/ labweb/ index. html). The expected distribution of nucleotide diversity and Tajima's *D* was generated by coalescent simulations using the software ms (Hudson 2002) with appropriate demographic model (Voight et al. 2005) or empirically using the resequencing data from Seattle SNP database (http:// pga.gs.washington.edu/). DNA divergence, Fu and Li's *D* and *F* (Fu and Li 1993) and its significance were calculated by DnaSP 5.10 (Librado and Rozas 2009). Haplotypes were inferred using the program PHASE 2.0 (Stephens and Donnelly 2003) and the phylogeny of

all haplotypes was reconstructed by the software NETWORK 4.5 (http:// www. fluxus-engineering. com) with median-joining algorithm (Bandelt et al. 1999). The time to the most recent common ancestor (MRCA) for these haplotypes was calculated by the formula, $T_{HC} \times D_H/D_{HC}$, in which T_{HC}, D_H, and D_{HC} indicated the divergence time between chimpanzee and human (~6 million years ago), the average distance between human haplotypes, and the divergence between the chimpanzee and human sequences (74.4 for resequenced region), respectively. Since we could not find the homologous region for segment 1 in the chimpanzee genome, this segment was omitted from phylogeny reconstruction and MRCA time estimate. The HapMap genotyping data (http:// hapmap. org) and *UGT2B4* resequencing data from the Environmental Genome Project (EGP, http:// egp. gs. washington. edu/ welcome. html) were also incorporated in some analysis. Tag SNPs were chosen by ldSelect (Carlson et al. 2004) with $r^2 \geq 0.8$.

Tag SNP genotyping in breast cancer case–control populations

Two hundred and sixty, 138, and 863 breast cancer patients and 102, 330, and 393 women without breast cancer were enrolled from the United States, Barbados, and Nigeria, respectively. All individuals were of African ancestry. Tag SNP rs13129471 was genotyped in these populations by a custom Taqman assay (Applied Biosystems) including primer pair 5′-TGGACTCATCACCTGACTCATGTAA-3′ and 5′-GTCAAAGAGACTGCAGGAACATGA-3′ and probe pair 5′-VIC-ATGCACACTATTCTGAAATA-3′ and 5′-FAM-ATGCACACTATTTTGAAATA-3′ (target site underlined) according to manufacturer's suggestion. Hardy–Weinberg equilibrium (HWE) was tested using a Chi-square test with one degree of freedom in cases and controls for three populations separately. The genotype frequencies in patients and controls were compared by Chi-square tests, followed by logistic regression model to estimate odds ratios (OR) and 95% confidence intervals (CI). Because of allele frequency difference across populations, the logistic regression adjusted an indicator variable for population. All statistical tests were performed in Stata 11.0 (StataCorp LP, College Station, TX, USA) and a $P < 0.05$ was considered statistically significant.

Results and discussion

Within the 4.8 kb resequenced region, 56 SNPs were identified (see Table S3) and most of them were present in all three populations (Fig. 1). Consequently, a similar estimate of the population mutation rate parameter θ_w (~2.2 × 10−3) was obtained for all three population samples (Table 1). Consistent with the HapMap data, most SNPs were in near-perfect LD (result not shown) and occurred at intermediate (~30–50%) minor allele frequency (MAF; see Table S3) in all three populations. As a result, a considerably high nucleotide diversity (Table 1), ~4.2 per kb in YRI and CEU and 3.5 in ASN, was observed; this value is ~3.7–4.5 times higher than the genome-wide average (~0.75 per kb; Reich et al. 2002; Sachidanandam et al. 2001) or in chromosome 4 (~0.81 per kb; Sachidanandam et al. 2001).

Under evolutionary neutrality, the vast majority of variants are rare. However, in the region upstream of *UGT2B4*, most variation occurs at intermediate frequency in all three populations (see Table S3). This pattern resulted in a strongly positive Tajima's *D* value in

YRI (2.6311, P = 0.0009), CEU (3.514, P < 0.0001), and ASN (1.6896, *P* = 0.055, see Table 1). We also compared the *UGT2B4* upstream region to 322 resequenced human genes from the Seattle SNP database to assess whether the pattern observed at *UGT2B4* is unusual relative to a large collection of human loci. For the YRI and CEU populations, the Tajima's *D* value for the *UGT2B4* upstream region is higher than the value for all genes in African Americans and European Americans (result not shown), respectively, thus confirming that the pattern in the *UGT2B4* upstream region is unusual in the human genome and is not likely to result from demographic history alone. It might be argued that this test might be biased since Seattle SNP data include some coding regions, which might be subjected to different selective pressures. However, considering the low proportion of coding region in resequenced region, the influence on distribution from coding region might be negligible.

Additional neutrality tests based on the frequency spectrum are Fu and Li's *D* and *F* (Fu and Li 1993), which compare external and internal mutations in a coalescent tree (Fu 1997). Here, we used the chimpanzee sequence as an outgroup and excluded segment 1 due to lack of a homologous segment in the chimpanzee genome. As shown in Table 1, all three populations showed a significantly positive Fu and Li's *D* and *F*, which indicate an excess of old mutations. This result further strengthens the proposal that the variation pattern in the *UGT2B4* upstream region departs from neutral expectations.

Between-population differentiation of allele frequency, which can be summarized by the F_{ST} statistic, can be enhanced by geographically restricted selective pressures, but may be decreased by balancing selection if selective pressures have similar intensity across populations (Akey et al. 2002). In the *UGT2B4* case, the F_{ST} values for most (92.9%) SNPs (see Table S3) are lower than the average value, 0.123, for non-coding region in the human genome among the three populations (Akey et al. 2002) and show a low degree of allele frequency differentiation overall.

We then used PHASE to infer the haplotype phase for the resequenced region. In total, 26 haplotypes were identified (data not shown). We also reconstructed the phylogeny for these haplotypes. As shown in Fig. 2, two major clusters were separated by as many as 30 mutation events. The two major haplotypes were found in all three populations at similar frequency while the rarer haplotypes tend to be specific to different populations (Fig. 2). The average pairwise sequence divergence between these clusters was 16.2, thus implying a divergence time of ~1.3 million years. This date is later than the divergence of human and chimpanzee, but precedes the dispersal of modern humans out of Africa (~80–100 thousand years ago; Cavalli-Sforza and Feldman 2003), which is not unusual but consistent with the presence of the two major haplotypes in the three major ethnic groups.

To determine whether the patterns of variation observed upstream of *UGT2B4* are unique to human, we also resequenced the homologous region in ten chimpanzees. None of the 56 sites that were polymorphic in human were shared with chimpanzee (see Fig. S3), consistent with the idea that these haplotypes originated after human–chimpanzee divergence. Moreover, unlike the patterns observed in humans, we detected relatively low diversity levels (π = 0.001066), and a slight excess of rare variants (Tajima's D = −0.2721).

Therefore, this region experienced a distinct evolutionary history between human and chimpanzee.

The striking pattern of frequency spectrum and the presence of two deep branches in the genealogy of the *UGT2B4* upstream region raise the possibility that balancing selection acted on this genomic region. The combined analysis of EGP *UGT2B4* resequencing and HapMap genotyping data indicates that the strong LD pattern starts at two promoter SNPs, rs941389 (−115 relative to translation start) and rs13129471 (−497), and spans 70 kb, ending at SNP rs7657504 (Fig. S1). Beyond this interval, the strong LD pattern decays dramatically (Fig. S1). This pattern of LD decay is more pronounced in the YRI population sample (results not shown), which is consistent with many previous studies showing lower LD levels in populations of African ancestry (The International HapMap Consortium 2007; Wang et al. 2006). This region includes numerous cis-regulatory elements of *UGT2B4* expression that have been validated through functional in vitro assays, such as binding sites for HNF1 homeobox A (HNF1A), POU class 2 homeobox 1 (POU2F1; Gardner-Stephen and Mackenzie 2008), peroxisome proliferatoractivated receptor α (PPARα; Barbier et al. 2003a), and farnesoid X receptor (FXR; Barbier et al. 2003b; Zhou et al. 2005). Therefore, it is possible that the two haplotypes have different promoter and/or enhancer activity and that the polymorphism underlying the signature of balancing selection influences the spatial or temporal regulation of *UGT2B4* expression. However, due to the length of the pattern, it was difficult to identify the target of selection based on available data.

Steroid hormones, including estrogen, androgen, glucocorticoid, mineralocorticoid, and progestine, are crucial for many aspects of human physiology, especially in initiating the development and maintaining the function of human reproductive system, anti-inflammatory reaction, response to stress, maintaining salt and water balance, and mediating menstrual cycle and pregnancy (Paul and John 2007). Therefore, deviations from proper steroid hormone levels, whether deficiency or excess, are likely to be deleterious (White 1994). UGT2B gene expression plays an important role in steroid hormone metabolism. For example, when the UGT2B17 gene is not expressed because it is removed by a common polymorphic deletion, the steroid hormone clearance is dramatically influenced (Jakobsson et al. 2006; Juul et al. 2009). Although the relationship between *UGT2B4* expression and steroid hormone metabolism has not been investigated in detail, it is plausible that alterations in *UGT2B4* expression levels have profound consequences on steroid hormone levels in human body. Based on these considerations, we propose that natural selection acted on the fine tuning of *UGT2B4* gene expression in order to keep an optimal level of this protein and, as a consequence, of the steroid hormones metabolized by *UGT2B4*. In this regard, it is interesting to note that a signal of balancing selection was also proposed for the UGT2B17 gene, though its functional significance has not been fully investigated (Xue et al. 2008).

Although balancing selection could provide an interpretation for the observed deviation from neutrality in the frequency spectrum, we cannot rule out the possibility that this pattern results from a partial selection sweep in human populations. Polymorphisms tightly linked to an advantageous mutation driven to intermediate frequency by selection will also tend to occur at intermediate frequency, thus inducing a skew in the allele frequency spectrum

(Biswas and Akey 2006; Nielsen et al. 2007), as observed above. Moreover, since this process is rapid, little decay of linkage disequilibrium is expected to occur, thus resulting in a signal in haplotype structure detectable by test statistics such as the extended haplotype homozygosity (EHH; Sabeti et al. 2002) or the integrated haplotype score (iHS; Voight et al. 2006). Although no significant deviation from neutral expectations was observed for either EHH or iHS in this region (result not shown), this possibility might still persist since the power of these tests is incomplete (Voight et al. 2006).

If, indeed, the *UGT2B4* upstream region contains a regulatory polymorphism, this variant is expected to have some phenotypic effects. Given the important role of steroid hormone levels in the susceptibility to breast cancer, we hypothesized that the two *UGT2B4* haplotypes might result in different breast cancer predisposition. To investigate this possibility, we genotyped tag SNP rs13129471 in three different case–control populations with African ancestry. No deviation from HWE was observed ($P > 0.05$) in all three populations (result not shown). The G allele frequency was 0.363 in healthy Nigerian women, compared with 0.578 in African Americans and 0.612 in Barbadians. This allele frequency difference suggests that stratified analysis or pooled analysis adjusting for population structure is warranted. After adjusting for population origin, the OR for heterozygous genotype and homozygous genotype for the G allele was 1.02 (95% CI 0.81–1.28) and 1.50 (95% CI 1.15–1.95), respectively, when compared with homozygous genotype for the A allele ($P = 0.003$; Table 2). The association was highly significant in Nigerians ($P = 0.000017$), with heterozygote OR = 1.16 and homozygote OR = 2.38 (Table 2). However, only weak association was observed in African Americans and no association was observed in Barbadians (Table 2), which might be due to relatively small sample sizes or incomplete correction for population structure in these two samples. Table 3 shows demographics and environmental risk factor information for our breast cancer patients. We examined whether these environment risk factors modify the effect of SNP rs13129471 and found no effect modification. These results suggest that these two haplotypes can influence breast cancer susceptibility and G allele of rs13129471 is the risk allele in Nigerian women. Moreover, considering that high steroid hormone level is a risk factor for breast cancer (Guillemette et al. 2004; Thijssen and Blankenstein 1989), we proposed that G allele of rs13129471 is associated with lower *UGT2B4* expression.

Acknowledgments

We thank Prof. Yoav Gilad and Dr. Jin-Xian Liu (University of California at Irvine) for providing chimpanzees genomic DNA and technical advice, respectively. We also thank the anonymous reviewers for their helpful comments. This research was supported by National Institutes of Health (CA125183 and U01 GM61393 to A.D.R.).

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 2002; 12:1805–1814. [PubMed: 12466284]

Allerston CK, Shimizu M, Fujieda M, Shephard EA, Yamazaki H, Phillips IR. Molecular evolution and balancing selection in the flavin-containing monooxygenase 3 gene (FMO3). Pharmacogenet Genomics. 2007; 17:827–839. [PubMed: 17885620]

Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, Clark AG, Nielsen R. Targets of balancing selection in the human genome. Mol Biol Evol. 2009; 26:2755–2764. [PubMed: 19713326]

Andres AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin S-Q, Hurle B, Schwartzberg PL, Williamson SH, Bustamante CD, Nielsen R, Clark AG, Green ED. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genet. 2010; 6:e1001157. [PubMed: 20976248]

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK. A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. Proc Natl Acad Sci USA. 2002; 99:10539–10544. [PubMed: 12149450]

Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999; 16:37–48. [PubMed: 10331250]

Barbier O, Duran-Sandoval D, Pineda-Torra I, Kosykh V, Fruchart JC, Staels B. Peroxisome proliferator-activated receptor alpha induces hepatic expression of the human bile acid glucuronidating UDP-glucuronosyltransferase 2B4 enzyme. J Biol Chem. 2003a; 278:32852–32860. [PubMed: 12810707]

Barbier O, Torra IP, Sirvent A, Claudel T, Blanquart C, Duran-Sandoval D, Kuipers F, Kosykh V, Fruchart JC, Staels B. FXR induces the UGT2B4 enzyme in hepatocytes: a potential mechanism of negative feedback control of FXR activity. Gastroenterology. 2003b; 124:1926–1940. [PubMed: 12806625]

Belanger A, Pelletier G, Labrie F, Barbier O, Chouinard S. Inactivation of androgens by UDP-glucuronosyltransferase enzymes in humans. Trends Endocrinol Metab. 2003; 14:473–479. [PubMed: 14643063]

Biswas S, Akey JM. Genomic insights into positive selection. Trends Genet. 2006; 22:437–446. [PubMed: 16808986]

Black FL, Hedrick PW. Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. Proc Natl Acad Sci USA. 1997; 94:12452–12456. [PubMed: 9356470]

Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R, Green P, Olson MV. Scan of human genome reveals no new Loci under ancient balancing selection. Genetics. 2006; 173:2165–2177. [PubMed: 16751668]

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M. The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol. 2008; 9:R143. [PubMed: 18817538]

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Bresolin N, Sironi M. Genetic variability in the ACE gene region surrounding the Alu I/D polymorphism is maintained by balancing selection in human populations. Pharmacogenet Genomics. 2010a; 20:131–134. [PubMed: 20038859]

Cagliani R, Riva S, Biasin M, Fumagalli M, Pozzoli U, Lo Caputo S, Mazzotta F, Piacentini L, Bresolin N, Clerici M, Sironi M. Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection. Hum Mol Genet. 2010b; 19:4705–4714. [PubMed: 20843824]

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; 74:106–120. [PubMed: 14681826]

Cauley JA, Lucas FL, Kuller LH, Stone K, Browner W, Cummings SR. Elevated serum estradiol and testosterone concentrations are associated with a high risk for breast cancer Study of Osteoporotic Fractures Research Group. Ann Intern Med. 1999; 130:270–277. [PubMed: 10068384]

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003; 33(Suppl):266–275. [PubMed: 12610536]

Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006; 2:e64. [PubMed: 16683038]

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Langaney A, Excoffier L. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. Am J Hum Genet. 2002; 70:207–223. [PubMed: 11741197]

Desai AA, Innocenti F, Ratain MJ. UGT pharmacogenomics: implications for cancer risk and cancer therapeutics. Pharmacogenetics. 2003; 13:517–523. [PubMed: 12893990]

Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997; 147:915–925. [PubMed: 9335623]

Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993; 133:693–709. [PubMed: 8454210]

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. A population genetics study of the familial Mediterranean fever gene: evidence of balancing selection under an overdominance regime. Genes Immun. 2009a; 10:678–686. [PubMed: 19675583]

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. 2009b; 19:199–212. [PubMed: 18997004]

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Riva S, Clerici M, Bresolin N, Sironi M. Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med. 2009c; 206:1395–1408. [PubMed: 19468064]

Gardner-Stephen DA, Mackenzie PI. Liver-enriched transcription factors and their role in regulating UDP glucuronosyltransferase gene expression. Curr Drug Metab. 2008; 9:439–452. [PubMed: 18537579]

Grigorova M, Rull K, Laan M. Haplotype structure of FSHB, the beta-subunit gene for fertility-associated follicle-stimulating hormone: possible influence of balancing selection. Ann Hum Genet. 2007; 71:18–28. [PubMed: 17227474]

Guillemette C. Pharmacogenomics of human UDP-glucuronosyltransferase enzymes. Pharmacogenomics J. 2003; 3:136–158. [PubMed: 12815363]

Guillemette C, Belanger A, Lepine J. Metabolic inactivation of estrogens in breast tissue by UDP-glucuronosyltransferase enzymes: an overview. Breast Cancer Res. 2004; 6:246–254. [PubMed: 15535854]

Hirayasu K, Ohashi J, Kashiwase K, Takanashi M, Satake M, Tokunaga K, Yabe T. Long-term persistence of both functional and non-functional alleles at the leukocyte immunoglobulin-like receptor A3 (LILRA3) locus suggests balancing selection. Hum Genet. 2006; 119:436–443. [PubMed: 16501917]

Hollox EJ, Armour JA. Directional and balancing selection in human beta-defensins. BMC Evol Biol. 2008; 8:113. [PubMed: 18416833]

Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–338. [PubMed: 11847089]

Izukawa T, Nakajima M, Fujiwara R, Yamanaka H, Fukami T, Takamiya M, Aoki Y, Ikushiro S, Sakaki T, Yokoi T. Quantitative analysis of UDP-glucuronosyltransferase (UGT) 1A and UGT2B expression levels in human livers. Drug Metab Dispos. 2009; 37:1759–1768. [PubMed: 19439486]

Jakobsson J, Ekstrom L, Inotsume N, Garle M, Lorentzon M, Ohlsson C, Roh HK, Carlstrom K, Rane A. Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. J Clin Endocrinol Metab. 2006; 91:687–693. [PubMed: 16332934]

Juul A, Sorensen K, Aksglaede L, Garn I, Rajpert-De Meyts E, Hullstein I, Hemmersbach P, Ottesen AM. A common deletion in the uridine diphosphate glucuronyltransferase (UGT) 2B17 gene is a strong determinant of androgen excretion in healthy pubertal boys. J Clin Endocrinol Metab. 2009; 94:1005–1011. [PubMed: 19088161]

Key TJ. Serum oestradiol and breast cancer risk. Endocr Relat Cancer. 1999; 6:175–180. [PubMed: 10731106]

King CD, Rios GR, Green MD, Tephly TR. UDP-glucuronosyltransferases. Curr Drug Metab. 2000; 1:143–161. [PubMed: 11465080]

Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25:1451–1452. [PubMed: 19346325]

Liu X, Fu Y, Liu Z, Lin B, Xie Y, Liu Y, Xu Y, Lin J, Fan X, Dong M, Zeng K, Wu CI, Xu A. An ancient balanced polymorphism in a regulatory region of human major histocompatibility complex

is retained in Chinese minorities but lost worldwide. Am J Hum Genet. 2006; 78:393–400. [PubMed: 16465617]

Mackenzie PI, Bock KW, Burchell B, Guillemette C, Ikushiro S, Iyanagi T, Miners JO, Owens IS, Nebert DW. Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. Pharmacogenet Genomics. 2005; 15:677–685. [PubMed: 16141793]

Nagar S, Remmel RP. Uridine diphosphoglucuronosyltransferase pharmacogenetics and cancer. Oncogene. 2006; 25:1659–1672. [PubMed: 16550166]

Nakamura A, Nakajima M, Yamanaka H, Fujiwara R, Yokoi T. Expression of UGT1A and UGT2B mRNA in human normal tissues and various cell lines. Drug Metab Dispos. 2008; 36:1461–1464. [PubMed: 18480185]

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007; 8:857–868. [PubMed: 17943193]

Norman P, Cook M, Carey BS, Carrington CF, Verity D, Hameed K, Ramdath DD, Chandanayingyong D, Leppert M, Stephens HF, Vaughan RW. SNP haplotypes and allele frequencies show evidence for disruptive and balancing selection in the human leukocyte receptor complex. Immunogenetics. 2004; 56

Ohno S, Nakajin S. Determination of mRNA expression of human UDP-glucuronosyltransferases and application for localization in various human tissues by real-time reverse transcriptase-polymerase chain reaction. Drug Metab Dispos. 2009; 37:32–40. [PubMed: 18838504]

Paul, W.; John, BD. Introduction to endocrinology. In: Gardner, DG.; Shoback, D., editors. Greenspan's basic and clinical endocrinology. 8th edn. New York: The McGraw-Hill Companies; 2007. p. 1-34.

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet. 2002; 32:135–142. [PubMed: 12161752]

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419:832–837. [PubMed: 12397357]

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409:928–933. [PubMed: 11237013]

Soejima M, Tachida H, Tsuneoka M, Takenaka O, Kimura H, Koda Y. Nucleotide sequence analyses of human complement 6 (C6) gene suggest balancing selection. Ann Hum Genet. 2005; 69:239–252. [PubMed: 15845028]

Sparks R, Ulrich CM, Bigler J, Tworoger SS, Yasui Y, Rajan KB, Porter P, Stanczyk FZ, Ballard-Barbash R, Yuan X, Lin MG, McVarish L, Aiello EJ, McTiernan A. UDP-glucuronosyltransferase and sulfotransferase polymorphisms, sex hormone concentrations, and tumor receptor status in breast cancer patients. Breast Cancer Res. 2004; 6:R488–R498. [PubMed: 15318931]

Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 2003; 73:1162–1169. [PubMed: 14574645]

Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA. Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet. 2006; 38:375–381. [PubMed: 16493422]

Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983; 105:437–460. [PubMed: 6628982]

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]

Tan Z, Shon AM, Ober C. Evidence of balancing selection at the HLA-G promoter region. Hum Mol Genet. 2005; 14:3619–3628. [PubMed: 16236759]

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]

Thijssen JH, Blankenstein MA. Endogenous oestrogens and androgens in normal and malignant endometrial and mammary tissues. Eur J Cancer Clin Oncol. 1989; 25:1953–1959. [PubMed: 2632276]

Thomas HV, Reeves GK, Key TJ. Endogenous estrogen and postmenopausal breast cancer: a quantitative review. Cancer Causes Control. 1997; 8:922–928. [PubMed: 9427435]

Tukey RH, Strassburg CP. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. Annu Rev Pharmacol Toxicol. 2000; 40:581–616. [PubMed: 10836148]

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet. 2002; 71:1112–1128. [PubMed: 12378426]

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA. 2005; 102:18508–18513. [PubMed: 16352722]

Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006; 4:e72. [PubMed: 16494531]

Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Nat Acad Sci. 2006; 103:135–140. [PubMed: 16371466]

Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975; 7:256–276. [PubMed: 1145509]

Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984; 38:1358–1370.

White PC. Disorders of aldosterone biosynthesis and action. N Engl J Med. 1994; 331:250–258. [PubMed: 8015573]

Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet. 2004; 74:637–646. [PubMed: 14997422]

Wright S. Genetical structure of populations. Nature. 1950; 166:247–249. [PubMed: 15439261]

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Huang N, Zerjal T, Lee C, Carter NP, Hurles ME, Tyler-Smith C. Adaptive evolution of UGT2B17 copy-number variation. Am J Hum Genet. 2008; 83:337–346. [PubMed: 18760392]

Zhou J, Zhang J, Xie W. Xenobiotic nuclear receptor-mediated regulation of UDP-glucuronosyl-transferases. Curr Drug Metab. 2005; 6:289–298. [PubMed: 16101569]

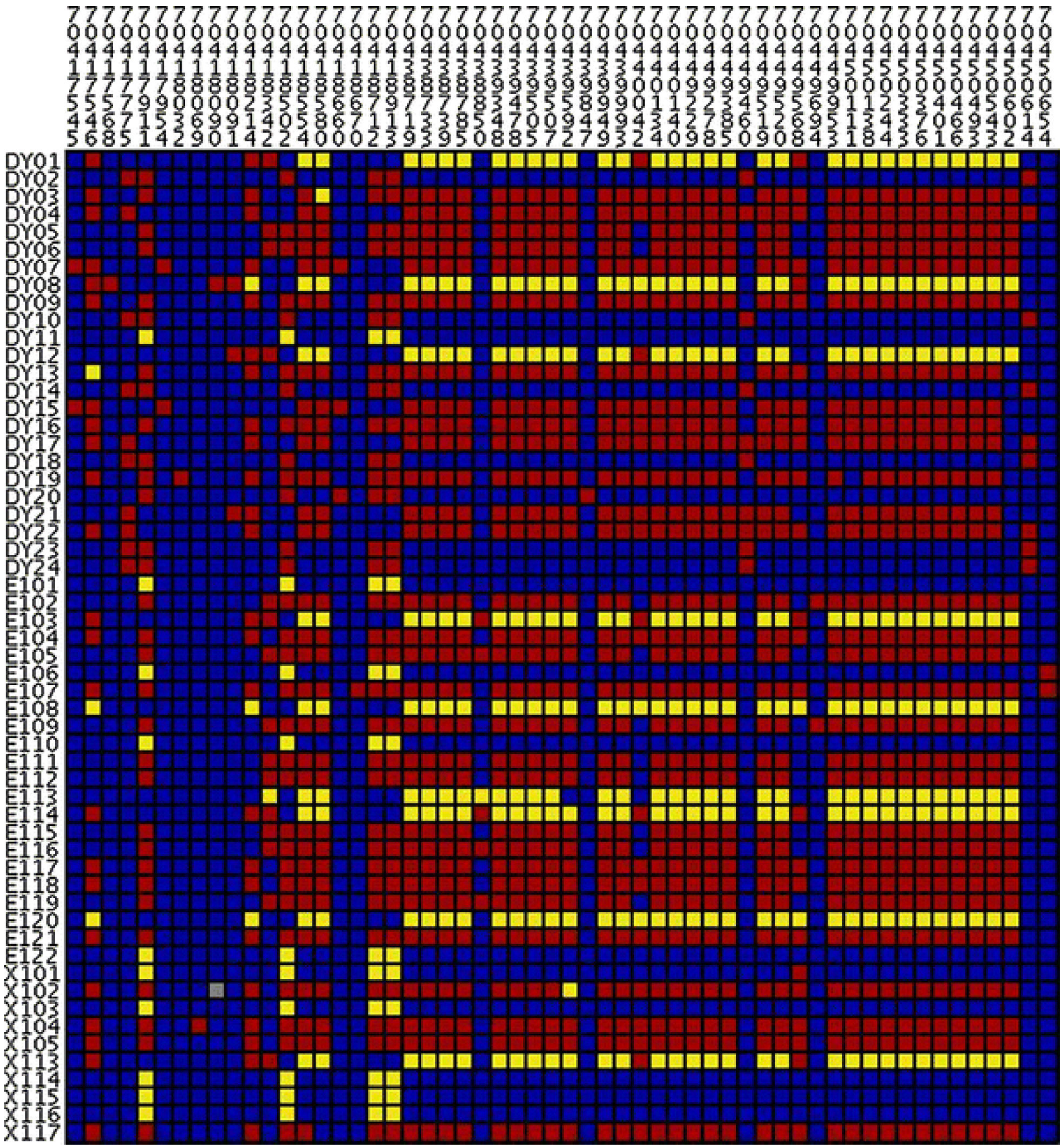

Fig. 1.

Visual genotype of UGT2B4 upstream region. Each column indicates one SNP while each row denotes one individual. Blue, red, yellow, and gray represent homozygous of common allele, heterozygous, homozygous of rare allele, and missing data, respectively. DY, E, and X indicate YRI, CEU, and ASN HapMap populations, respectively. All positions refer to the genome sequence (build 36) for chromosome 4.

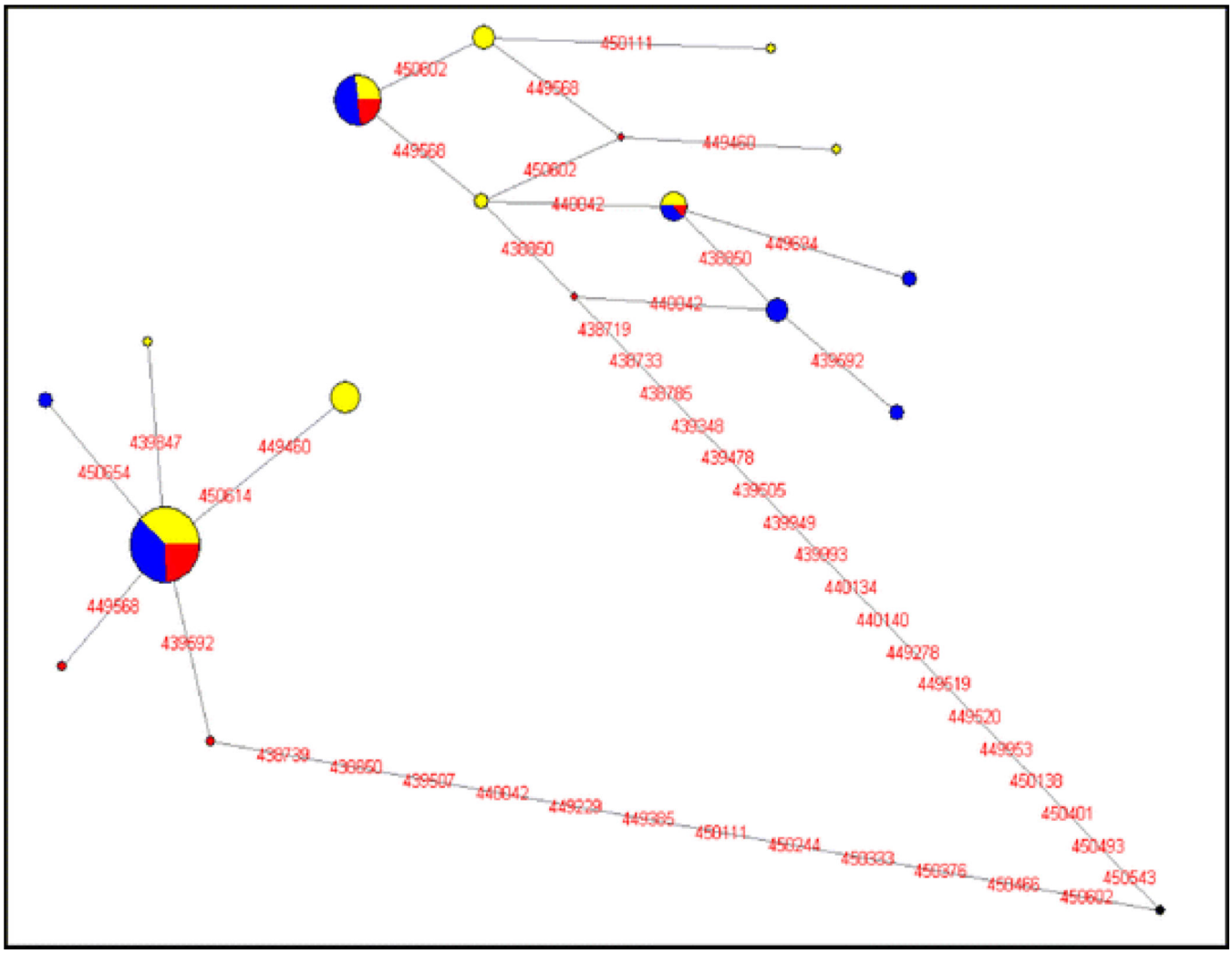

Fig. 2.

Phylogeny of the haplotypes in UGT2B4 upstream region. Each circle represents a unique haplotype and its area was proportional to the frequency of the haplotype. Within each node, YRI, CEU, and ASN are labeled in red, blue, and yellow, respectively. The positions of the mutations differentiating the haplotypes are shown on each branch (all position should add 70000000; build 36). The chimpanzee sequence was taken as outgroup.

Table 1

Summary statistics for the resequenced regions in human *UGT2B4* upstream region

	n [a]	***S*** [b]	θ_w	π	**Tajima'*D* (*P*)[c]**	**Fu and Li's *D* (*P*)**	**Fu and Li's *F*(*P*)**
YRI	48	51	0.002406	0.004228	2.6311 (0.0009)	1.70908 (<0.05)	2.83113 (<0.02)
CEU	44	44	0.002118	0.004252	3.514 (<0.0001)	1.99802 (<0.02)	3.08716 (<0.02)
ASN	20	40	0.002479	0.003528	1.6896 (0.055)	1.66768 (<0.05)	2.15988 (<0.02)

[a] Number of chromosomes

[b] Number of segregating sites

[c] *P* value is from simulation and presented in parentheses

Table 2

Genotype distribution of rs13129471 in women with breast cancer and healthy controls

Population	Genotype	Cases (%)	Controls (%)	P	OR (95% CI)
African Americans	A/A	54 (20.8)	14 (13.7)	0.029	1.00 (ref.)
	A/G	108 (41.5)	58 (56.9)		0.48 (0.25–0.94)
	G/G	98 (37.7)	30 (29.4)		0.85 (0.41–1.73)
Barbadians	A/A	27 (19.6)	54 (16.4)	0.55	1.00 (ref.)
	A/G	64 (46.4)	148 (44.8)		0.86 (0.50–1.49)
	G/G	47 (34.0)	128 (38.8)		0.73 (0.42–1.30)
Nigerians	A/A	274 (31.7)	155 (39.4)	1.66E−5	1.00 (ref.)
	A/G	391 (45.3)	191 (48.6)		1.16 (0.89–1.51)
	G/G	198 (22.9)	47 (12.0)		2.38 (1.64–3.46)
Total	A/A	355 (28.2)	223 (27.0)	0.003	1.00
	A/G	563 (44.6)	397 (48.2)		1.02 (0.81–1.28)
	G/G	343 (27.2)	205 (24.8)		1.50 (1.15–1.95)

Table 3

Risk factors in breast cancer cases and controls tested for SNP rs13129471

Risk factor	**Cases (*n* = 1,261)**	**Controls (*n* = 825)**	***P***
Age, mean ± SD	47.4 ± 12.1	46.3 ± 15.5	0.075
Menopausal status, *n* (%)			
Premenopausal	612 (53.4)	460 (57.9)	0.05
Postmenopausal	534 (46.6)	334 (42.1)	
Positive family history of breast cancer, *n* (%)	167 (14.8)	43 (5.5)	<0.001
Body mass index (kg/m^2), mean ± SD	26.5 ± 6.2	27.3 ± 6.4	0.007
Height (cm), mean ± SD	160.9 ± 7.2	160.1 ± 6.9	0.01
Parity, mean ± SD	3.7 ± 2.5	3.0 ± 2.5	<0.001
Age at first live birth, mean ± SD	22.8 ± 5.0	22.5 ± 5.0	0.19
Oral contraceptives, *n* (%)	344 (34.2)	140 (30.2)	0.13