Genomic evidence for recent positive selection at the human MDR1 gene locus

Kun Tang, Li Peng Wong, Edmund J.D. Lee, Samuel S. Chong and Caroline G.L. Lee

1Department of Biochemistry, 2Department of Pharmacology and 3Department of Pediatrics and Obstetrics/Gynecology, National University of Singapore, Singapore, 4Division of Medical Sciences, National Cancer Center, Singapore and 5Department of Pediatrics and Gynecology/Obstetrics, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

Received August 21, 2003; Revised and Accepted February 10, 2004

The MDR1 multidrug transporter regulates the traffic of drugs, peptides and xenobiotics into the body as well as sensitive tissues like the brain, germ cells and the developing fetus. Hence, it may influence an individual’s response to drugs as well as his/her susceptibility to complex diseases in which environmental factors, especially xenobiotics, play a role. Polymorphisms within this gene, especially single-nucleotide polymorphism e26/3435(C/T), have been variously associated with differences in MDR1 expression, function, drug response and disease susceptibility. Here, we report the detailed characterization of the haplotype and linkage disequilibrium architecture of the entire 200 kb of the MDR1 gene in five world populations, namely, Chinese, Malays, Indians, Caucasians and African-Americans. We observed varied haplotype diversity across the entire gene in the different populations. The major haplotype mh5, which contains the subhaplotype e12/1236T–e21/2677T–e26/3435T, is highly represented among the four non-African populations, while mh7, which contains the subhaplotype e12/1236C–e21/2677G–e26/3435C, accounts for over a third of African-American chromosomes. These observations are inconsistent with a simple population evolution model, but instead are suggestive of recent historical events that have maintained such long range linkage disequilibrium. Using a modified long-range haplotype test, we found statistically significant evidence of recent positive selection for the e21/2677T and e26/3435T alleles in the Chinese population, and for the e26/3435T allele in the Malay population. Interestingly, we also detected evidence for positive selection of the alternative allele e26/3435C in the African-American population. These data suggest that independent mutational events may have occurred on the mh5 and mh7 haplotypes of the MDR1 gene to confer positive selection in the non-African and African-American populations, respectively.

INTRODUCTION

The MDR1 multidrug transporter is one of the better characterized of the ATP-binding cassette (ABC) family of transporters. Although this transporter was initially identified in the context of its contribution to the phenomenon of multidrug resistance (MDR) against anticancer drugs (1), its range of drugs and substrates as well as its implied physiological role has greatly expanded. In addition to chemotherapeutic drugs, the MDR1 transporter has been found to efflux a wide variety of substrates representing nearly every category of clinically important drugs including anti-arrhythmics, anti-depressants, anti-psychotics and anti-virals (2–4). This transporter can also transport peptides across membranes (5,6) and has been implicated in the regulation of entry of enveloped viruses (e.g. HIV-1 and influenza virus) into cells (7,8). Mouse knockout (9) as well as tissue and cell-specific expression studies highlight the importance of the MDR1 transporter in protecting the body against xenobiotic insults and drug toxicity at two levels. The first level of protection involves the regulation of the absorption of xenobiotics/drugs into the body via MDR1 expression in the epithelial cells of the gut while the
second level of protection involves the regulation of traffic of xenobiotics/drugs into sensitive tissues like the brain (10), germ-cells or fetus (11) via its expression at the blood–brain and blood–germ cell/fetal barriers, respectively. Hence, the MDR1 transporter not only influences drug response and pharmacological outcome, it may also influence an individual’s susceptibility to viral infection or to complex disorders in which environmental factors, particularly xenobiotics, play a role.

The MDR1 gene spans 200 kb and is located at chromosome 7p21. It comprises 28 exons and possesses two transcription start sites and two promoters that are ~100 kb apart. In the past few years, great interest has been generated in the identification and characterization of single nucleotide polymorphisms (SNPs) in the MDR1 gene and its association with drug response, clinical outcome and susceptibility to certain diseases. These studies examined SNPs from the downstream promoter to the 3′ end of the gene and primarily concentrated on SNP e26/3435(C/T), as well as SNPs e2/677(G/T/A), e12/1236(C/T) and e1–129(T/C). Except for SNP e1–129(T/C), low frequency (<10%) exonic SNPs have been reported in only one or two ethnic groups and may represent ethnic-specific or geographically restricted SNPs. Of the high-frequency exonic SNPs, only three SNPs, e12/1236(C/T), e21/2677(G/T/A) and e26/3435(C/T), have been reported to occur in different ethnic populations and at different frequencies (12).

The high but ethnically variable frequency SNP e26/3435(C/T) has been variously associated, either alone or in combination with SNPs e21/2677(G/T/A), e12/1236(C/T) or e1–129(T/C), with differences in MDR1 expression (13–21) and function (13–16, 18, 19, 22–29), drug response and clinical outcome (16, 21, 23, 30–33), and with susceptibility to various diseases (20, 21, 34–37). SNP e26/3435(C/T), however, does not result in an amino acid change and conflicting results of positive, negative and no associations have been reported. We previously demonstrated strong but varied linkage disequilibrium (LD) of SNP e26/3435(C/T) with SNPs e21/2677(G/T/A) and e12/1236(C/T) in three different Asian populations, and proposed that SNP e26/3435(C/T) may not be the causal modulator of previously reported functional differences but may instead be in strong linkage disequilibrium with different unidentified causal SNPs in the different study populations (12). Resolution of the various contradictory association reports may benefit from detailed knowledge of the genetic structure and evolutionary history of the entire MDR1 gene and the ability to detect disease association.

We have now characterized the haplotype and LD profiles across the entire 200 kb of the MDR1 gene including both the distal and proximal promoters in the same three Asian populations as well as in the Caucasian and African-American populations. Consistent with the earlier study, we found significant variation in both SNP allele and haplotype frequencies amongst the five different populations. We also observed low haplotype diversity across the entire 200 kb of the MDR1 gene in all five ethnic groups, with one haplotype accounting for between 25 and 50% of all chromosomes analyzed. We also present genomic evidence of recent positive selection for different alleles of SNPs e21/2677 and e26/3435 in the African-American and non-African populations.

RESULTS

MDR1 SNP allele frequencies differ among populations

In a previous study, we discovered that only four of 10 SNPs in the promoter and coding regions of the MDR1 gene were polymorphic in the Chinese, Malay and Indian populations represented in Singapore. However, strong LD was observed between all three high-frequency coding SNPs, which span 40 kb of the 3′ half of the gene (12). Here, we analyzed these four SNPs, together with eight additional SNPs, in these three populations as well as in Caucasians and African-Americans.

As shown in Table 1, the six SNPs downstream of the proximal promoter (SNPs 7–12) have high minor allele frequencies (>14%) in all five populations, while the six SNPs spanning the region upstream of the proximal promoter (SNPs 1–6) have lower minor allele frequencies, with some SNPs (SNPs 2, 4 and 5) being monomorphic in one or more ethnic groups. Re-sequencing of 1.7 kb of the proximal promoter in 140 chromosomes of mixed ethnic origin did not identify any new SNPs of high frequency (data not shown). Other studies which also screened for SNPs in the proximal promoter and 5′–UTR also failed to detect high-frequency SNPs (17, 38, 39), suggesting that the upstream region of the MDR1 gene has a lower sequence diversity than the rest of the gene.

None of the 12 SNPs examined deviated significantly from Hardy–Weinberg equilibrium although allele frequencies of several SNPs varied significantly among the different ethnic groups (Table 1). Overall, different ethnic groups showed similar allele frequencies for some SNPs but not others. For example, allele frequencies for SNPs 3, 4 and 12 are similar between Chinese and Malays, and between Indians, Caucasians and African-Americans. SNP 4 may be a population-specific SNP because the SNP 4G allele accounts for 9.78 and 4.89% of Chinese and Malay chromosomes, respectively, but represents only 0.5% of Caucasian and African-American alleles and is absent in the Indians (P < 0.005; Table 1). Strikingly in the African-Americans, allele frequencies for several high-frequency SNPs, namely 8, 9, 10 and 11, were distinctly different from the non-African populations (P < 0.001; Table 1). SNPs 2 and 5 were excluded from further analyses as they were monomorphic in four of the five populations studied.

MDR1 haplotype diversity differs among populations

Haplotype frequencies of 10 SNPs, excluding SNPs 2 and 5, were derived from genotype data using the expectation–maximization (EM) algorithm. These 10 SNPs were included in our determination of haplotype frequency for two reasons. Firstly, although several of the SNPs occur at low frequency or were monomorphic in one or two population groups, these SNPs were found not to significantly affect the EM estimation of haplotype frequencies at other loci. Secondly, many of the low frequency or population-specific SNPs occur in potentially functional regions like the promoter; hence it was of interest to determine which haplotype these SNP alleles were assigned to.

Assuming random association among the 10 SNPs, a simulation utilizing over 1000 iterations based on the tested sample size and observed allele frequencies (40) predicted
Table 1. Allele frequency comparisons of the different SNPs in the different populations

<table>
<thead>
<tr>
<th>SNP no.</th>
<th>SNP ID</th>
<th>JSNP IDa</th>
<th>Region</th>
<th>Population</th>
<th>n</th>
<th>HWE P-value</th>
<th>Allele frequency (%)</th>
<th>Pairwise differences Fisher's exact P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>e1/-195</td>
<td>ssj0000001</td>
<td>Upstream promoter</td>
<td>CH</td>
<td>96</td>
<td>0.59</td>
<td>C 10.42 T 89.58</td>
<td>0.05 0.05 0.06 0.73</td>
</tr>
<tr>
<td>2</td>
<td>i1/-40423</td>
<td>ssj0000003</td>
<td>Intron -1</td>
<td>CH</td>
<td>96</td>
<td>1.00</td>
<td>A 100.00 C 0.00</td>
<td>1.00 1.00 1.00 1.00</td>
</tr>
<tr>
<td>3</td>
<td>i1/-3655</td>
<td>ssj0000007</td>
<td>Intron -1</td>
<td>CH</td>
<td>96</td>
<td>1.00</td>
<td>A 85.94 G 14.06</td>
<td>0.22 0.15 0.21 5.90 × 10^-4</td>
</tr>
<tr>
<td>4</td>
<td>i1/-41</td>
<td>ssj0000008</td>
<td>Intron 1</td>
<td>CH</td>
<td>92</td>
<td>0.19</td>
<td>A 90.22 G 9.78</td>
<td>0.11 1.83 × 10^-6 1.49 × 10^-5 1.49 × 10^-5</td>
</tr>
<tr>
<td>5</td>
<td>e1/-145</td>
<td>NA^c</td>
<td>Exon 1</td>
<td>CH</td>
<td>92</td>
<td>1.00</td>
<td>A 99.46 G 0.54</td>
<td>1.00 0.49 0.48 0.48</td>
</tr>
<tr>
<td>6</td>
<td>e1/-129</td>
<td>ssj0000009</td>
<td>Exon 1</td>
<td>CH</td>
<td>92</td>
<td>1.00</td>
<td>A 4.35 G 95.65</td>
<td>1.00 0.06 0.81 0.21</td>
</tr>
<tr>
<td>7</td>
<td>i3/-5890</td>
<td>MS-JST 146287</td>
<td>Intron 3</td>
<td>CH</td>
<td>96</td>
<td>0.66</td>
<td>A 61.98 G 38.02</td>
<td>0.25 0.07 0.26 1.93 × 10^-2</td>
</tr>
<tr>
<td>8</td>
<td>i6/-293</td>
<td>ssj0000014</td>
<td>Intron 6</td>
<td>CH</td>
<td>96</td>
<td>0.51</td>
<td>A 39.06 G 60.94</td>
<td>0.75 0.24 2.42 × 10^-3 1.85 × 10^-11</td>
</tr>
<tr>
<td>9</td>
<td>e12/1236</td>
<td>ssj0000017</td>
<td>Exon 12</td>
<td>CH</td>
<td>92</td>
<td>0.08</td>
<td>A 38.04 G 61.96</td>
<td>1.00 0.52 2.10 × 10^-3 9.20 × 10^-18</td>
</tr>
<tr>
<td>10</td>
<td>e21/2677</td>
<td>ssj0007866</td>
<td>Exon 21</td>
<td>CH</td>
<td>92</td>
<td>0.19</td>
<td>A 13.04 G 86.96</td>
<td>0.50 5.36 × 10^-6 3.74 × 10^-3 2.65 × 10^-10</td>
</tr>
</tbody>
</table>

Continued
112.31 ± 3.33, 101.85 ± 3.28, 69.23 ± 2.48, 84.11 ± 2.58 and 59.99 ± 2.63 different haplotypes in the Chinese, Malay, Indians, Caucasian and African-Americans, respectively. However, only 11 haplotypes could be observed if neither recombination nor recurrent mutation occurred (40). In this study, we identified 48, 43, 33, 38 and 39 different haplotypes in the Chinese, Malay, Indian, Caucasian and African-American populations, respectively (Fig. 1). Lower haplotype diversity observed in Indians, Caucasians and African-Americans was probably due to the fact that more SNPs were monomorphic in Indians (three), Caucasians and African-Americans (two each) compared to Chinese and Malays (one each). The observation of lower haplotype diversity in the African-Americans appears to contradict the notion that African-Americans would have a higher haplotype diversity on account of their older population age. One possible explanation for the observed low haplotype diversity, especially in the African-American population, could be attributed to ascertainment biases in the choice of SNPs, many of which were primarily selected from the JSNP public database.

Table 1. Continued

<table>
<thead>
<tr>
<th>SNP no.</th>
<th>SNP ID</th>
<th>JSNP ID*</th>
<th>Region</th>
<th>Population</th>
<th>n</th>
<th>HWE P-value</th>
<th>Allele frequency (%)</th>
<th>Pairwise differences Fisher's exact P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>T</td>
<td></td>
<td></td>
<td>CH</td>
</tr>
<tr>
<td>11</td>
<td>e26/3435</td>
<td>ssj0000047 Exon 26</td>
<td>CH</td>
<td>92</td>
<td>0.27</td>
<td>C</td>
<td>58.15</td>
<td>41.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ML</td>
<td>92</td>
<td>0.83</td>
<td>T</td>
<td>59.24</td>
<td>40.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IN</td>
<td>95</td>
<td>0.82</td>
<td></td>
<td>37.37</td>
<td>62.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CAU</td>
<td>100</td>
<td>0.42</td>
<td></td>
<td>44.50</td>
<td>55.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AA</td>
<td>100</td>
<td>0.13</td>
<td></td>
<td>79.00</td>
<td>21.00</td>
</tr>
<tr>
<td>12</td>
<td>e28/4036</td>
<td>ssj0000052 Exon 28</td>
<td>CH</td>
<td>92</td>
<td>1.00</td>
<td>A</td>
<td>75.54</td>
<td>24.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ML</td>
<td>92</td>
<td>1.00</td>
<td>G</td>
<td>76.09</td>
<td>23.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IN</td>
<td>95</td>
<td>1.00</td>
<td></td>
<td>85.26</td>
<td>14.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CAU</td>
<td>100</td>
<td>1.00</td>
<td></td>
<td>86.50</td>
<td>13.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AA</td>
<td>100</td>
<td>1.00</td>
<td></td>
<td>89.00</td>
<td>11.00</td>
</tr>
</tbody>
</table>

*Japanese Single Nucleotide Polymorphism Database Identifier.

At position –1 (Fig. 1), we observed the lowest haplotype diversity in Afri- can-Americans (two each) compared to Chinese and Malays (one each). The latter two populations groups share an additional number of major haplotypes (mh5, mh6 and mh7; Fig. 2IV), which also carried the least significant difference between the Indians and the Caucasians (P = 0.00084) and highly significant difference for all other pair-wise comparisons between populations (P < 0.00001).

Highly variable LD between SNP loci

We determined pair-wise LD between SNPs using |D| and r², which are useful for modeling recombination rates and association power, respectively (41). LD generally decreased with physical distance, although large variation was observed in the LD-distance relationship (Fig. 2B). Data points in the plot of |D| versus physical distance were highly scattered, with some physically close SNP pairs showing nearly undetectable LD (e.g. |D| = 0.097 between SNPs 11 and 12 in the Indians; Fig. 2IA), while some SNP pairs that are separated by ~100 kb showed complete LD (e.g. |D| = 1 between SNPs 3 and 12 in the Chinese; Fig. 2IA). Although a genetically older population such as the African-Americans would be expected to have generally lower LD, this was not obvious in our pair-wise LD analyses for the entire MDR1 gene, as the LD0.5 determined from |D| versus distance trend lines for African-Americans (~110 kb) was not much shorter compared to the non-African populations (~110–150 kb; Fig. 2B). The best correlation between |D| and physical distance was observed in the Indian population (r² = 0.1191), which also carried the least number of major haplotypes (mh5, mh6 and mh7; Fig. 2IBB), highlighting the effect of underlying haplotype structure on local LD patterns.

Overall, useful LD was found to extend over shorter distances when r² was utilized as a measure of LD. Utilizing r² ≥ 0.3 as the threshold for useful LD for association studies (41,42), we observed that useful LD spans the shortest distance...
Figure 1. Haplotype profiles of the 10 MDR1 SNPs in the five populations. Haplotype frequencies were derived from genotype data of the 10 SNPs (excluding SNP 2 and 5) using the expectation–maximization algorithm. In all, 118 different haplotypes occur in at least one of the five populations, of which 48, 43, 33, 38 and 39 haplotypes were found in the Chinese, Malay, Indian, Caucasian and African-American populations, respectively. Haplotype profiles are presented as horizontal arrays of black, white or gray squares. Each column represents a SNP locus, labeled with the corresponding SNP ID. Each row represents a haplotype, with the estimated haplotype frequencies expressed as percentage shown in the adjacent table as well as portrayed as a bar graph. The 13 most frequent haplotypes, with frequencies above 4% in at least one population, are labeled mh1–mh13, respectively.
Figure 2. Pairwise linkage disequilibrium profiles for single SNPs (A) $D^\prime$, $r^2$ as well as Fisher's exact $P$-value of association were calculated for each pair of the 10 SNPs in the five populations (I, Chinese; II, Malays; III, Indians; IV, Caucasian; V, African-American) and summarized into tables. Cells above the diagonal indicate $D^\prime$ values, whereas those below the diagonal contain the $r^2$ values. Gradient of either red ($D^\prime$) or blue ($r^2$) color for each cell is presented according to the scale specified beside the tables, with darker colors representing higher $D^\prime$ or $r^2$. Pairs of SNPs that have significant linkage disequilibrium as determined by the Fisher's exact test are indicated with asterisks, with $P$-value between 0.01 and 0.05 specified by a single-asterisk; $P$-value between 0.001 and 0.01 specified by a double-asterisk and $P$-values below 0.001 specified by a triple-asterisk. ‘—’ indicates pairs where calculations of LD cannot be applied. (B) $D^\prime$ between SNP pairs was calculated (see Materials and Methods) and plotted against their physical distances in the five populations. A classical LD decay model was fitted to derive the LD decay trend line (see Materials and Methods). The $r^2$ coefficients are also indicated. (C) $r^2$ between SNP pairs was calculated (see Materials and Methods) and plotted against their physical distances in the five populations.
Figure 3. HBDs for five selected loci in the five populations. Haplotype branching diagrams are presented for five selected loci: SNPs 1 and 12, representing the 5’ end and 3’ end, respectively, as well as SNPs 9, 10 and 11 which showed visual evidence of possible positive selection in at least one population. Diagrams for each population were aligned in columns and each row of diagrams has a different tested root locus. The breakdown of LD on each allele of the tested locus is portrayed bidirectionally in a parallel manner. The alleles of the tested locus, denoted by black dots with the alleles labeled, serve as roots of the LD-breakdown trees. The position of each locus along the x-axis is scaled to the physical distance between each locus. Each locus is denoted as a node specified by a black dot. Lines are drawn between nodes to represent the specific haplotypes. The thickness of the lines as well as size of the black dots corresponds to the frequency of the haplotype. Depending on the presence of possible extended haplotypes, branches may or may not be seen.
in the Chinese and Malays and was longest in the Indians (Fig. 2A and C). The higher \( r^2 \) values and longer useful LD distance in the Indian and Caucasian populations reflect the predominance of only two major haplotypes, mh5 and mh7, in Indians (49.1 and 15.7%, respectively) and Caucasians (33.8% and 18.8%, respectively; Fig. 1). This is because \( r^2 \) is a measure of ’perfect linkage disequilibrium’ and achieves a value of 1 when only two out of four possible gamete types are present for two bi-allelic markers.

### SNPs 10 and 11 of the MDR1 gene are positively selected

Recently, a powerful test for recent positive selection, called the long-range haplotype (LRH) test, was proposed by Sabeti et al. (43). This test is based on the relationship between the frequency of an allele and the extent of LD surrounding the allele. The primary rationale of this test is that under the assumption of neutral evolution, recent variants should occur at low frequency in a population and require extended periods before reaching high frequency. Over extended periods, allele frequency increases, LD surrounding this variant allele will decay substantially due to recombination. As such, high-frequency variants are typically more ancient and their surrounding LD extends only over short distances. However, recent variants with extensive long-range LD and high allele frequency cannot be accounted for by neutral genetic drift, but are instead indicative of recent positive selection. One major advantage of the LRH test is that it is not sensitive to population-specific or local recombination rates, by utilizing alternative allele values as internal controls to adjust for variations in local recombination rate among different populations (43).

In this study, we adapted the LRH test for analysis on single SNPs instead of core haplotypes, and modified the algorithm to handle multiallelic SNPs. Instead of haplotype bifurcation diagrams, bidirectional haplotype branching diagrams (HBD) of the 12 SNPs were generated for all five populations (see Materials and Methods). Selected HBDs with roots representing SNPs at the proximal and distal ends of the MDR1 gene as well as internal SNPs showing potential positive selection are presented in Figure 3. As illustrated in the HBDs, the Chinese and Malays share similar numbers and complexity of branching whereas the Indians resemble the Caucasians in HBD profile, with fewer branches indicating that Indians and Caucasians are less diverse in this gene. On the contrary, the African-Americans differed greatly from the non-African populations in their overall haplotype branching patterns (Fig. 3).

Consistent with a neutral evolution hypothesis, we observed the majority of high-frequency root alleles produced greater numbers of branchings compared with lower frequency root alleles. Inconsistent with the assumption of neutral evolution, however, some root alleles produced only a single predominant branch, such as alleles SNP 10T and SNP 11T in the four non-African populations. These variants occur at high frequency in these populations, yet maintained a single predominant haplotype across most of the 200 kb region, as represented by the single thick branch in each HBD profile. The main branch corresponds to the most common major haplotype, mh5, shared amongst the four non-African populations (Fig. 3), unlike their alternative alleles, which have two or more main branches. In contrast, mh5 occurs at much lower frequency in African-Americans, while mh7, which carries alleles SNP 10G and SNP 11C, accounts for 34% of chromosomes in this group (Figs 1 and 3).

When the extended haplotype homozygosity (EHH) of SNPs 10 and 11 was plotted against distance for all the five populations, the EHH of SNP variants 10T and 11T decayed more slowly than their corresponding control/alternative alleles, SNP 10A or G and SNP 11C, in the four non-African populations (Fig. 4, left panel). In the Chinese, the EHH of SNP variants 11T and 10T over ∼200 kb were 0.60 and 0.68, respectively, whereas the EHH of their control/alternative alleles 11C and 10A or G over the same distance were between 0.13 and 0.20 (Fig. 4, left top panel). No obvious difference was observed in EHH decay of either SNP variants 11T or 10T and their corresponding alternative alleles in the African-Americans (Fig. 4, left bottom panel).

To test for evidence of recent positive selection at the MDR1 gene locus, we plotted the relative EHH against their allele frequencies (Fig. 5, middle and right panels) using a modified method proposed by Sabeti et al. (43). Relative EHH is defined as the ratio of the EHH of the tested allele relative to the EHH of the control/alternative allele(s) (see Materials and Methods). \( P \)-values were generated by comparing the relative EHH against the simulated data at the observed allele frequencies. SNPs 10 and 11 were selected for these analyses since these variants, although occurring at high frequency in these populations, maintained a single predominant haplotype across most of the 200 kb region, as represented by the single thick branch in each HBD profile (Fig. 3). Furthermore, both these SNPs have been associated with various functional phenotypes (13–16,18–29,32–37).

Representative relative EHH plots are shown in the middle and right panels of Figure 4 and the \( P \)-values are summarized in Table 2. In the Chinese population, the relative EHH of the T-allele of SNPs 10 and 11 are 4.8 and 4.7, respectively (Fig. 4, middle and right panels). Both these alleles demonstrate significant departure from evolutionary neutrality under four different population model assumptions and three different recombination rate assumptions \((P < 0.05; Table 2)\). Significant evidence of positive selection under all tested population model and recombination rate assumptions was also observed in the Malays, but only for the T-allele of SNP 10 \((P < 0.05)\). In the Indians and Caucasians, evidence of positive selection of either SNP variants 10T or 11T achieved statistical significance only under some population model and recombination rate assumptions. Interestingly and in marked contrast to the findings in the Chinese and Malay populations, the African-American population showed statistically significant evidence of positive selection for the alternative C-allele of SNP11 (Table 2).

### DISCUSSION

Numerous studies have associated the high frequency synonymous SNP 11 with differences in MDR1 expression and function, plasma drug concentration, drug-induced side effects, drug response as well as disease susceptibility. Interestingly, allele frequencies of this SNP vary greatly among different
Figure 4. EHH and relative EHH tests. Left column: the EHH at varying distances from the SNPs 10 and 11 was plotted for the Chinese, Malay, Indian, Caucasian and African-American populations. The EHH decay line for SNP10A in African-Americans was omitted as this allele is present in only 1/200 African-American chromosomes. Middle and right columns: relative EHH at the most distant SNP (see Materials and Methods) was plotted against allele frequency for each allele of SNPs 10 and 11 (black diamond), and compared against simulated data (gray dots) under the constant-size population model (see Materials and Methods). The simulations in this figure assume a recombination rate of 1.3 cM Mb⁻¹, except for SNP11 in Indians and Caucasians denoted by an asterisk. Simulation data for these two populations at a recombination rate of 0.65 cM Mb⁻¹ were shown for these two plots. The 95th, 75th and 50th percentile lines of the distribution of simulated alleles were also given.
populations (12). However, contradictory observations of positive, negative and no associations with this allele have been reported. It was hypothesized that strong LD between this SNP and different unobserved causal SNP in the different study population may account for the conflicting reports (12). Consideration of the evolutionary history and genetic variation of the MDR1 gene in each population may aid in understanding these contradictions. For example, population changes or recent positive/negative selection of specific alleles can influence haplotype frequencies, linkage disequilibrium, and ultimately the likelihood of detecting an association.

Varied haplotype diversity and long-range LD in the MDR1 gene

In this study, we found that the overall profile of haplotype distribution based on haplotype frequencies is most similar between the Chinese and Malays ($P = 0.116$). Some similarity in the haplotype distribution was observed between the Indian and Caucasian populations ($P = 0.00084$), but the African-Americans differed the most from the non-African populations ($P < 0.00001$; Fig. 1). This trend was also observed schematically in the haplotype branching diagrams (Fig. 3). Consistent with the 'Out-of-Africa' hypothesis whereby all modern populations originated from a common ancestral population in Africa (44–46), all the major haplotypes represented in the non-African populations, except for mh6 and mh13 (Fig. 1), were also found in the African-American population. Overall, our data are consistent with the hypothesis of an ancient divergence of the ancestor of the non-African populations from the ancestral African population, followed by an early divergence of this non-African ancestor into the common ancestors of Caucasians/Indians and Chinese/Malays, a later divergence between the Caucasians and the Indians, and most recently, a divergence between the Malays and Chinese. This hypothesis is consistent with the original geographical distribution of these five populations.

Notably, we found varied haplotype diversity in the entire 200 kb of the MDR1 gene (Fig. 1). Out of an expected 60–112 haplotypes from the 10 SNPs under the assumption of linkage equilibrium, we observed that between three and six haplotypes alone could account for more than 60% of total chromosomes within each population, with the lowest number of major haplotypes observed in the Indians (three), followed by the Caucasians (four), Chinese (five), and Malays and African-Americans (six each; Fig. 1).

LD between SNP pairs was found to vary considerably and did not always exhibit an inverse relationship with distance (Fig. 2). Within each population, low LD between physically close SNPs as well as high LD between distant SNPs could be observed. Such highly variable LD profiles have also been reported for other gene regions in various populations (41,47–50). Interestingly, the LD0.5 at this gene locus for the African-American population is much longer (~110 kb) than those observed by Reich et al. (51) for other gene loci (~5 kb). Factors that may contribute to the observed exceptionally longer LD at this gene locus could be the choice of SNPs for our analyses, possible ‘recombination coldspot’ around the MDR1 gene locus, or a recent historical event such as severe bottleneck or positive selection for this gene.

Figure 5. Distribution of 12 SNPs across the MDR1 gene. The horizontal line represents the entire length of the MDR1 gene, and individual exons are represented by vertical bars above the line. The twelve SNPs that we examined are represented as vertical bars beneath the horizontal line, labeled with the systematic names for each SNP given. The nomenclature for these SNPs are as follows: region/position number (allele types), where region denotes whether it is in an exon (e) or intron (i); position number is either the mRNA nucleotide position using the translation start site as position 1 if the SNP is in an exon, or the position in the genomic sequence using the 5' boundary of the immediate downstream exon as position $-1$. Ue-1/-195(T/C) denotes that the SNP is located upstream of exon $-1$ at the $-195$th position. This map was drawn to scale according to the scale bar.
Evidence of recent positive selection at the MDR1 gene locus

Interestingly, the mh5 haplotype (Fig. 1), which contains the sub-haplotype 9T–10T–11T, occurs at highest frequency in the four non-African populations, with frequencies of 31.3, 25.5, 49.1 and 33.8% in the Chinese, Malays, Indians and Caucasians, respectively, while present in only 5.3% of African-American chromosomes. Interestingly, although the mh5 haplotype occurs at a very low frequency in the African-American population, the mh7 haplotype, which contains the mh5 haplotype occurs at very low frequency in the African-American chromosomes. Interestingly, although the mh5 haplotype (Fig. 1), which contains the alternative C-allele. These observations suggest that two independent mutational events may have occurred on different allele/haplotype backgrounds to confer selective advantage on the SNP 10T/11T allele(s) of the mh5 haplotype and the SNP 11C allele of the mh7 haplotype in the non-African and African-American populations, respectively. To further evaluate this hypothesis, the LRH test should be conducted on additional populations using a panel of SNPs that is informative in all the tested populations.

In the Indians and Caucasians, statistically significant P-values were observed for SNP 10 or 11 only under one or two population and/or recombination rate assumptions. The failure to unequivocally detect positive selection in the Indians and Caucasians could be real or due to the masking effect of a severe population bottleneck or of a structured population, or due to the choice of Asian-oriented SNPs used in this study. Thus it may be advantageous to identify more informative high frequency SNPs to facilitate better evaluation of the EHH profile in all populations.

Implication of recent positive selection with respect to functional disease association studies

Our observation of possible recent positive selection on the SNP 10 and 11 alleles is interesting in the light of the many African-American population also showed statistically significant evidence of positive selection for SNP 11, but for the alternative C-allele. These observations suggest that two independent mutational events may have occurred on different allele/haplotype backgrounds to confer selective advantage on the SNP 10T/11T allele(s) of the mh5 haplotype and the SNP 11C allele of the mh7 haplotype in the non-African and African-American populations, respectively. To further evaluate this hypothesis, the LRH test should be conducted on additional populations using a panel of SNPs that is informative in all the tested populations.

In the Indians and Caucasians, statistically significant P-values were observed for SNP 10 or 11 only under one or two population and/or recombination rate assumptions. The failure to unequivocally detect positive selection in the Indians and Caucasians could be real or due to the masking effect of a severe population bottleneck or of a structured population, or due to the choice of Asian-oriented SNPs used in this study. Thus it may be advantageous to identify more informative high frequency SNPs to facilitate better evaluation of the EHH profile in all populations.

Implication of recent positive selection with respect to functional disease association studies

Our observation of possible recent positive selection on the SNP 10 and 11 alleles is interesting in the light of the many African-American population also showed statistically significant evidence of positive selection for SNP 11, but for the alternative C-allele. These observations suggest that two independent mutational events may have occurred on different allele/haplotype backgrounds to confer selective advantage on the SNP 10T/11T allele(s) of the mh5 haplotype and the SNP 11C allele of the mh7 haplotype in the non-African and African-American populations, respectively. To further evaluate this hypothesis, the LRH test should be conducted on additional populations using a panel of SNPs that is informative in all the tested populations.

Evidence of recent positive selection at the MDR1 gene locus

Interestingly, the mh5 haplotype (Fig. 1), which contains the sub-haplotype 9T–10T–11T, occurs at highest frequency in the four non-African populations, with frequencies of 31.3, 25.5, 49.1 and 33.8% in the Chinese, Malays, Indians and Caucasians, respectively, while present in only 5.3% of African-American chromosomes. Interestingly, although the mh5 haplotype occurs at a very low frequency in the African-American population, the mh7 haplotype, which contains the mh5 haplotype occurs at very low frequency in the African-American chromosomes. Interestingly, although the mh5 haplotype (Fig. 1), which contains the alternative C-allele. These observations suggest that two independent mutational events may have occurred on different allele/haplotype backgrounds to confer selective advantage on the SNP 10T/11T allele(s) of the mh5 haplotype and the SNP 11C allele of the mh7 haplotype in the non-African and African-American populations, respectively. To further evaluate this hypothesis, the LRH test should be conducted on additional populations using a panel of SNPs that is informative in all the tested populations.

In the Indians and Caucasians, statistically significant P-values were observed for SNP 10 or 11 only under one or two population and/or recombination rate assumptions. The failure to unequivocally detect positive selection in the Indians and Caucasians could be real or due to the masking effect of a severe population bottleneck or of a structured population, or due to the choice of Asian-oriented SNPs used in this study. Thus it may be advantageous to identify more informative high frequency SNPs to facilitate better evaluation of the EHH profile in all populations.

Implication of recent positive selection with respect to functional disease association studies

Our observation of possible recent positive selection on the SNP 10 and 11 alleles is interesting in the light of the many

Evidence of recent positive selection at the MDR1 gene locus

Interestingly, the mh5 haplotype (Fig. 1), which contains the sub-haplotype 9T–10T–11T, occurs at highest frequency in the four non-African populations, with frequencies of 31.3, 25.5, 49.1 and 33.8% in the Chinese, Malays, Indians and Caucasians, respectively, while present in only 5.3% of African-American chromosomes. Interestingly, although the mh5 haplotype occurs at a very low frequency in the African-American population, the mh7 haplotype, which contains the mh5 haplotype occurs at very low frequency in the African-American chromosomes. Interestingly, although the mh5 haplotype (Fig. 1), which contains the alternative C-allele. These observations suggest that two independent mutational events may have occurred on different allele/haplotype backgrounds to confer selective advantage on the SNP 10T/11T allele(s) of the mh5 haplotype and the SNP 11C allele of the mh7 haplotype in the non-African and African-American populations, respectively. To further evaluate this hypothesis, the LRH test should be conducted on additional populations using a panel of SNPs that is informative in all the tested populations.

In the Indians and Caucasians, statistically significant P-values were observed for SNP 10 or 11 only under one or two population and/or recombination rate assumptions. The failure to unequivocally detect positive selection in the Indians and Caucasians could be real or due to the masking effect of a severe population bottleneck or of a structured population, or due to the choice of Asian-oriented SNPs used in this study. Thus it may be advantageous to identify more informative high frequency SNPs to facilitate better evaluation of the EHH profile in all populations.

Implication of recent positive selection with respect to functional disease association studies

Our observation of possible recent positive selection on the SNP 10 and 11 alleles is interesting in the light of the many

Evidence of recent positive selection at the MDR1 gene locus

Interestingly, the mh5 haplotype (Fig. 1), which contains the sub-haplotype 9T–10T–11T, occurs at highest frequency in the four non-African populations, with frequencies of 31.3, 25.5, 49.1 and 33.8% in the Chinese, Malays, Indians and Caucasians, respectively, while present in only 5.3% of African-American chromosomes. Interestingly, although the mh5 haplotype occurs at a very low frequency in the African-American population, the mh7 haplotype, which contains the mh5 haplotype occurs at very low frequency in the African-American chromosomes. Interestingly, although the mh5 haplotype (Fig. 1), which contains the alternative C-allele. These observations suggest that two independent mutational events may have occurred on different allele/haplotype backgrounds to confer selective advantage on the SNP 10T/11T allele(s) of the mh5 haplotype and the SNP 11C allele of the mh7 haplotype in the non-African and African-American populations, respectively. To further evaluate this hypothesis, the LRH test should be conducted on additional populations using a panel of SNPs that is informative in all the tested populations.

In the Indians and Caucasians, statistically significant P-values were observed for SNP 10 or 11 only under one or two population and/or recombination rate assumptions. The failure to unequivocally detect positive selection in the Indians and Caucasians could be real or due to the masking effect of a severe population bottleneck or of a structured population, or due to the choice of Asian-oriented SNPs used in this study. Thus it may be advantageous to identify more informative high frequency SNPs to facilitate better evaluation of the EHH profile in all populations.

Implication of recent positive selection with respect to functional disease association studies

Our observation of possible recent positive selection on the SNP 10 and 11 alleles is interesting in the light of the many
reported associations of SNPs 11 and 10 with various diseases and with differences in MDR1 expression, function, plasma drug concentration and drug response and with certain diseases. Detailed analysis of the LD structure and distribution of the MDR1 haplotypes in each study population could clarify the basis of these associations, including the contradictory ones. Since SNPs 10 and 11 show considerable LD across the entire 200 kb MDR1 gene region (Fig. 2), strong association of either SNPs 10 or 11 with a disease would suggest that an unobserved causative variant could lie anywhere within the region defined by strong LD using the \( r^2 \) statistic, which has greater power to detect association between genetic markers and phenotype than the \( |D'| \) statistic (42). Strong \( r^2 \) LD was observed in the 90 kb between SNPs 7 and 12 in the Indians, in the 80 kb between SNPs 7 and 11 in the Caucasians and in the 60 kb between SNPs 8 and 11 in the Chinese and Malays (\( r^2 > 0.3, P < 0.01; \) Fig. 2). We recently tested several MDR1 SNPs and SNP haplotypes for evidence of association with Parkinson’s disease (PD) in the ethnic Chinese population (206 PD patients and 224 matched controls), and found that SNPs 10 (\( P = 0.00067 \)) and 11 (\( P = 0.00074 \)) as well as SNP haplotypes containing these two SNPs (\( P < 0.05 \)) are significantly associated with risk of developing PD (52). In particular, SNP alleles 10T and 11T were found to be significantly over-represented in the control population and under-represented in PD patients, suggesting that the T-allele of both SNPs may confer increased protection against xenobiotic insults at the blood–brain barrier, where the MDR1 transporter functions. Our present observation of strong genomic evidence of recent positive selection for SNP alleles 10T and 11T in the Chinese is remarkable when taken together with the results of the disease association study (52).

The MDR1 transporter is suggested to regulate the traffic of drugs, xenobiotics (4) and enveloped viruses (e.g. influenza virus) (7,8) across membranes, protecting the body as well as sensitive organs from xenobiotic insults. Our present study suggests that different selective forces may be acting at the MDR1 gene locus in African and non-African populations. Hence, it is important to understand the role of MDR1 in population evolution, as well as the effect of different MDR1 alleles on the fitness of their carriers. It is possible that drastic environmental changes like the recent viral epidemics, (e.g. the Spanish Flu of 1918) or the advent of the Industrial Revolution may have provided the necessary differential selective pressures that favored different alleles of the MDR1 gene in different populations over a relatively short period of human history. As the frequencies of these two SNPs vary among the different populations, it remains to be determined if they are also positively selected in other populations.

**MATERIALS AND METHODS**

**Study population**

DNA samples from 96 random, anonymized individuals from the three major ethnic Asian populations—Chinese, Indian and Malay—were obtained from discarded umbilical cord blood of neonates from the National University Hospital in Singapore. Except for their ethnicity, all original identifiers were removed and destroyed. Ethical approval for this study was obtained from the National University Hospital Institutional Review Board (NUH IRB). Genomic DNA for the Caucasian and African-American populations was obtained from the respective Human Variation Panels from the Coriell Cell Repositories (Camden, NJ, USA; http://locus.umdnj.edu/nigms/nigms cgi/panel.cgi?id=1).

**Selection of SNPs and genotyping**

A total of 12 SNPs were selected to span the entire ~200 kb of the MDR1 gene (Fig. 5). In addition to SNPs that occur in potentially functional regions like the proximal promoter, 5′ and 3′-UTR and coding regions as previously reported (12,53), additional SNPs in the introns and the distal promoter were included to generate a denser and more evenly distributed SNP map. These additional SNPs were selected from the JSNP public database (http://snp.imis.u-tokyo.ac.jp/) as SNPs from this database have been confirmed by sequencing. JSNP identifiers are listed for all the SNPs in this study (Table 1) except for SNP e1/-145(C/G), which we identified through sequencing but are not reported in any other public database. Each of the 12 SNPs are numbered from 1 to 12 as shown in Figure 5 and Table 1.

Multiplex PCR and minisequencing was employed for genotyping the 12 SNPs in two panels. The technique for multiplex PCR/minisequencing of seven of these SNPs has been described previously (53). A similar method was employed to genotype the remaining panel of five SNPs. Primer sequences, PCR and minisequencing conditions for the 12 SNPs in two panels. The technique for multiplex PCR/minisequencing of seven of these SNPs has been described previously (53). A similar method was employed to genotype the remaining panel of five SNPs. Primer sequences, PCR and minisequencing conditions for the genotyping of the five SNPs can be found in the Supplementary Material (Fig. S1).

**Statistical analyses: allele, haplotype frequency estimation and LD determination between SNP loci**

Fisher’s exact test was utilized to evaluate if allele frequencies of the various SNPs in each population deviate from Hardy–Weinberg equilibrium. Statistical significance of pair-wise comparisons of the allele frequencies between different populations was also evaluated using Fisher’s exact test. The population genetics analysis program Arlequin (http://www.arlequin.org/), was utilized to estimate haplotype frequencies based on the EM algorithm (54). Samples in which one or more loci cannot be genotyped due to reaction failure were excluded from haplotype frequency estimation. Fisher’s exact test was performed to compare haplotype frequencies between populations.

LD between SNP pairs was estimated using two statistics, namely the absolute Lewontin’s coefficient \( |D'| \) (42,55), and Pearson’s correlation \( r^2 \) (42,56). Statistical significance of LD between SNP pairs was assessed using Fisher’s exact test. To evaluate the overall LD in each population, the classical LD decay equation \( D = D_o(1 - \theta)^{j} \) (57) was utilized to obtain the LD decay trend line. The \( r^2 \) coefficients were also calculated to evaluate the extent this classical model can account for the observed \( |D'|–d\)istance relationship. Half-LD (LD\(_{0.5}\)) is defined as the distance at which \( |D'| \) is 0.5.
LRH test for recent positive selection

A modified LRH test (43) was adopted to test for positive selection on single SNPs rather than core haplotypes. In addition, instead of the HBD, which only allows the analysis of bi-allelic SNPs, a HBD algorithm was programmed using the VBA language in Microsoft Excel to enable the analysis of multi-allelic markers, including the tri-allelic SNP 10.

The EHH function of the LRH test was used to assess LD decay over distance. EHH at a distance $x$ from the tested variant (either an SNP allele or a core haplotype) is defined as the ratio of the EHH of the tested allele against the EHH of all other alleles at the same locus (43).

Coalescent simulations

Coalescent simulations (58) were performed to test the hypothesis of positive selection. Four different population models including constant population size model, expansion model, bottleneck model and structured population model were simulated. For each model, three different recombination rates, namely, $0.65, 1.3$ (59) and $2.6 \text{ cM Mb}^{-1}$ were simulated. Three mutation rates, $0.5 \times 10^{-8}$ site$^{-1}$ generation$^{-1}$, $1.0 \times 10^{-8}$ site$^{-1}$ generation$^{-1}$ and $2.0 \times 10^{-8}$ site$^{-1}$ generation$^{-1}$, were also tested for the constant-size model with recombination rate of $1.3 \text{ cM Mb}^{-1}$. The distribution of mutation sites are expected to be independent of genealogy, since these mutation sites were assigned after simulation of genealogy. The resulting data distribution was also found to be consistent despite different mutation rates (see Supplementary Material, Fig. S2). Thus, a constant mutation rate of $1.0 \times 10^{-8}$ site$^{-1}$ generation$^{-1}$ was utilized for the other models.

A sequence of 300 kb was simulated. Datasets matching the observed data to within $\pm 12.5\%$ of the allele frequency and EHH, and within $\pm 3\%$ in distance for all downstream loci were selected. The anchor locus was chosen to be within $\pm 3\%$ of the tested locus (either SNP 10 or 11) for the selected datasets. Each simulation was iterated at least 10 000 times to achieve $\sim 20 000–50 000$ data points for each of the five populations.

Plots of allele frequency versus relative EHH for the simulated data points were obtained and compared with the observed data. Probability lines of 95, 75 and 50% were obtained by binning the simulated data by allele frequency into 20 bins of equal size with intervals of 5%. P-values were computed by ranking the relative EHH of the observed SNP of interest with that of all of the simulated data points that lie within a $\pm 0.25$ allele frequency window of that SNP.

The parameters for the four tested models were defined as follows. A population size of 10 000 was assumed for the constant-sized model. The expansion model assumed a sudden population expansion from $10^4$ to $10^7$ occurring 200 generations ago. The bottleneck model assumed sudden population size shrinkage from 10 000 to 800 occurring 800 generations ago which recovered to 10 000 at the 640th generation. Two equal subpopulations with a constant size of 5000 and a constant migration rate $4N_{sub}m = 0.1$ was assumed for the structured population.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENT

This study was supported by a grant from the National Medical Research Council, Singapore (NMRC/0657/2002) to C.G.L.L., S.S.C and E.J.D.L.

REFERENCES


