Signatures of recent positive selection at the ATP-binding cassette drug transporter superfamily gene loci

Zihua Wang¹,³, Jingbo Wang¹, Erwin Tantoso⁴, Baoshuang Wang⁵, Amy Y.P. Tai⁵, London L.P.J. Ooi⁶,⁷, Samuel S. Chong²,⁸,* and Caroline G.L. Lee¹,⁵*

¹Department of Biochemistry, ²Department of Pediatrics, ³Graduate Programme in Bioengineering, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, ⁴Bioinformatics Institute, Singapore, ⁵Division of Medical Sciences, ⁶Department of Surgical Oncology, National Cancer Center, Singapore, ⁷Department of Surgery, Singapore General Hospital, Singapore and ⁸Children’s Medical Institute and Department of Laboratory Medicine, National University Hospital, Singapore

Received January 10, 2007; Revised March 19, 2007; Accepted March 31, 2007

Members of the ATP-binding cassette (ABC) superfamily of transporters have been implicated as major players in drug response. Single nucleotide polymorphisms (SNPs) in the ABC transporter genes may account for variation in drug response between individuals. Given the abundance of SNPs within the human genome, identification of functionally important SNPs is difficult. Here, we utilized signatures of recent positive selection (RPS) to identify SNPs in ABC genes that have potential functional significance by using the long-range-haplotype test to search for signatures of RPS at 18 ABC genes involved in drug transport. From the genotype data of these 18 ABC genes in four populations extracted from the HapMap database, at least one SNP in each of these genes displayed genomic signatures of RPS in at least one population. However, only 13 SNPs in 10 ABC genes from three populations retained statistical significance after Type I error reduction. The functional significance of six of these RPS SNPs, including those that failed multiple testing correction (MTC), has been reported previously. We experimentally confirmed a functional effect for two SNPs, including one that failed to show evidence of RPS after MTC. These observations suggest that Type I error reduction may inadvertently increase Type II error. Although the remaining positively selected SNPs have yet to be functionally validated, our study illustrates the feasibility of using this strategy to identify SNPs within ‘adaptive’ genes that may confer functional effect, prior to testing their roles in individual/population drug response variation or in complex disease susceptibility.

INTRODUCTION

Drug response can differ greatly amongst individuals. Different individuals may respond, respond partially, not respond or even experience adverse drug reactions (ADR) to the same dose of the same drugs. Differences in drug response can be attributed to the general health status of the individual, drug interactions or differences in age, sex etc. (1). In addition, genetic factors have been suggested to play an important role in the determination of a drug’s absorption, distribution, metabolism and interaction with its target. Genetic factors generally account for 15–30% of variability in drug metabolism and response, although for certain categories of drugs, they may account for as high as 95% of inter-individual differences (2–4). It is thus important to understand the genetic basis underlying inter-individual variation in drug response to facilitate the realization of personalized medicine, the ultimate goal of pharmacogenetics (5).

Drug response is a complex phenomenon involving the interaction of several different components including drug...
metabolizing enzymes, drug transporters, drug targets and regulatory factors. Recognized recently as key determinants of drug response is the ATP-binding cassette (ABC) group of efflux drug transporters such as MDR1 (6). The ABC superfamily is highly conserved throughout evolution and represents the largest class of transmembrane proteins. Currently, more than 48 members have been identified in this superfamily and they can be classified into seven distinct subfamilies, A to G, based on similarity in gene structure, amino acid sequence homology and domain organization (7). Amongst members of this superfamily are several members that are currently known to transport clinically relevant drugs, e.g. MDR1/ABCB1, ABCC1 through ABCC6 and ABCG2 (8).

Single nucleotide polymorphisms (SNPs) are the most abundant class of sequence variation in the human genome, accounting for ~90% of all inter-individual variation and representing one of the most common types of pharmacogenetic variation (9). Although the human genome harbors more than 10 million SNPs, accounting for >1% of the genome, only a fraction of all SNPs are functionally significant and account for differences between individuals (10).

With the increasing density of SNPs in the human genome, investigating each and every SNP individually for its functionality or association with disease/drug response becomes increasingly impractical, and identifying a functionally important SNP can be likened to finding a ‘needle-in-a-haystack’. Hence, several more cost-effective strategies have been adopted for genotype–phenotype association studies. One of these strategies exploits the redundancy of polymorphisms in regions of strong linkage disequilibrium (LD) to identify a minimal set of ‘tagging’ SNPs (tSNPs) to represent all SNPs in that gene (11–13). However, not all genes or genomic regions are amenable to this strategy as SNP-tagging efficiency is reduced in gene regions with low LD.

Another approach utilizes genomic signatures of recent positive selection (RPS) to identify regions that are, or have in the recent past been, functionally important (14). The rationale behind this approach is based on principles of natural selection and ‘survival of the fittest’. The exodus of mankind from Africa exposed our ancestors to new/different environments/diseases. Polymorphisms in genes facilitating better adaptation to the new environment would have conferred a survival advantage and rapidly increased in frequency, while deleterious polymorphisms would have been rapidly reduced in that population. The rapid increase or decrease in population frequency of polymorphisms under selective pressure, when compared with their flanking neutral polymorphisms, creates unique identifiable genomic ‘signatures’ that identify the selected SNPs in such ‘adaptive’ genes. With availability of the complete human genome sequence and increasing numbers of SNP variations deposited in various databases including the HapMap and Perlegen databases, several genetic algorithms have recently been developed to identify regions of the genome that are likely to be positively selected (15). Several large-scale genome-wide (16–18) and gene-specific studies (19,20) have recently been performed to identify genomic signatures of RPS.

We hypothesized that members of the ABC superfamily of drug transporter genes belong to the category of ‘adaptive’ genes that leave behind such genomic signatures. In this study, we employed the long-range-haplotype (LRH) test to search for signatures of RPS within 18 major ABC genes that have been implicated in drug transport (see Supplementary Material, Table S1), including ABCA2, ABCB1, ABCB4, ABCB11, ABCC1-C13 and ABCG2. The functional significance of two positively selected polymorphisms was further experimentally verified.

RESULTS

Significant differences in allele frequencies amongst the four populations at the ABCC11/ABCC12 gene loci

The 18 ABC transporter genes are located in 13 different chromosomes. Three of the genes (ABCB1, ABCB4 and ABCC7) reside on chromosome 7, of which ABCB1 and ABCB4 lie adjacent to each other on chromosome 7q21.1. Another four ABC genes are found on chromosome 16, of which ABCC1 and ABCC6 lie adjacent to one another on chromosome 16p13.1 and ABCC11 and ABCC12 reside together on chromosome 16q12.1. The adjacent locations of some of these genes suggest that they may have arisen through gene duplication.

A few observations can be made about these 18 ABC transporter genes based on the genotype data downloaded from the HapMap Project Release 16C.1 (June 2005). For each of these genes, which range in size from 18.67 to 281.59 kb, there were between 4 and 113 SNPs with genotype data available in all four populations that had a minor allele frequency (MAF) of ≥0.05 in at least one population (Table 1). Hence, the density of SNPs genotyped in this release of HapMap was approximately 1 SNP per 0.9–5.4 kb. Generally, greater than 45% of the genotyped SNPs in the ABC genes had a MAF ≥0.05 in at least one population (Table 1).

SNPs at the ABCC11 and ABCC12 gene loci, which reside together on chromosome 16q12.1, displayed an interesting trend. Although greater than 90% of SNPs had MAF ≥0.05, most of these high-frequency SNPs occurred in the YRI population alone (ABCC12) or YRI and CEU populations (ABCC11) (Table 1). Less than 7% of the SNPs in the CHB and JPT had MAF ≥0.05 suggesting potentially significant differences in allele frequencies amongst the different populations in these two adjacent genes (Table 1).

To evaluate the significance of these differences, $F_{ST}$ was calculated for all SNPs in the 18 ABC gene loci (Supplementary Material, Table S2). The mean $F_{ST}$ value for each gene is presented in Table 1. Highly significant differences were observed between the mean $F_{ST}$ values of the 41 SNPs at the ABCC11/ABCC12 gene region and the mean $F_{ST}$ values of the 530 SNPs at the other 16 ABC gene loci ($P < 0.001$) as well as the mean $F_{ST}$ values of 26 530 previously reported genome-wide SNPs ($P < 0.0017$) (21). Two SNPs, e5/G538A in ABCC11 and e24/C3349T in ABC12, had the two highest $F_{ST}$ values (0.81 and 0.68) out of a total of 571 SNPs from the 18 ABC genes. These data suggest that ABCC11 and ABCC12 may represent high-$F_{ST}$ candidate selection genes, although they were not among the 174 candidate selection genes previously reported (21).
Strong LD and low haplotype diversity within ABCB1, ABCB2, ABCC5 and ABCC7

When low frequency SNPs (MAF < 0.05) in the genic regions of the ABC genes were filtered out, there were insufficient SNPs for the determination of haplotype and LD profiles for ABCA2, ABCB10, ABCC11 and ABCC12 in all populations except for ABCC11 in the CEU and YRI populations and ABCC12 in the YRI (Table 1). For the remaining genes, the number of major haplotypes accounting for >60% of the total chromosomes per 100 kb and the half |D'| LD for the each gene in each population are shown in Table 1. Consistent with earlier studies (22–27), the YRI population displayed the weakest LD with average half LD of 67.6 kb when compared to >210 kb for the other populations. The YRI also showed the greatest haplotype diversity, with an average of 12 haplotypes per 100 kb accounting for >60% of the total chromosomes compared to <8 haplotypes for the other populations (Table 1).

Among the ABC genes, ABCB1, ABCB2, ABCC5 and ABCC7 generally displayed strong LD (>100 kb) and low haplotype diversity while ABCC1, ABCC3, ABCC8 and ABCC13 had comparatively weaker LD and higher haplotype diversity (Table 1). Similar observations were previously made for ABCB1 (25), ABCC5 (26) and ABCC1 (27).

Genomic evidence of RPS of various ABC genes in different populations

Table 2 and Figure 1 show SNPs within each ABC gene in each population that displayed genomic evidence of RPS. SNPs that continued to show evidence of RPS after correction for multiple testing are represented by multiple testing correction (MTC) in Table 2 and solid boxes in Figure 1, while SNPs that dropped out of significance after MTC are represented by LRH in Table 2 and dotted boxes in Figure 1.

Before Type I error reduction by correcting for multiple testing, at least one SNP in each ABC gene that we examined, with the exception of ABCC11 and ABCC4, displayed genomic evidence of RPS in one or two populations (Table 2). Interestingly, when the ABCC11 and ABCC4 SNPs were re-examined using more comprehensive genotype data from a newer release of the HapMap Project (release 20, Jan 06), SNPs i4/C4542T in ABCB4 and e24/A3084G in ABCC11 now yielded a positive result for RPS in the CEU when possible Type I error was not taken into consideration. Hence, a total of 32 SNPs were found to display RPS, most of them being selected only in a single population except for SNPs e35/T5562C in ABCB2, e21/G2677T/A in ABCC1, e9/C504T in ABCB4 and i1/T1541G and i4/A473C in ABCC11, which showed RPS in two populations (Table 2). This is consistent with the current algorithm employed, which detects only recent selection events that occur after population separation (17). It is possible for a SNP to be positively selected in two different populations if the selection event occurred just before the population divergence. This explanation appears unlikely for SNP e35/T5562C in ABCB2, which is positively selected in the YRI and CHB. Alternatively, two geographically separate populations could share a similar positively selected SNP if they were under similar selection events. Interestingly, the major T and A alleles of SNPs i1/T1541G and i4/A473C, respectively, in ABCC11 that showed RPS in both the YRI and CEU (Table 2) were found to be fixed in the CHB and JPT (Supplementary Material, Table S2). This could be because the selection event could be stronger or could have
Table 2. Summary of ABC SNPs showing genomic evidence of RPS

<table>
<thead>
<tr>
<th>Gene</th>
<th>Positively selected SNPs</th>
<th>dbSNP ID</th>
<th>allele</th>
<th>Modified LR test&lt;sup&gt;a&lt;/sup&gt;</th>
<th>iHS&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNP name</td>
<td>Location</td>
<td>AA change</td>
<td>CHB</td>
<td>JPT</td>
</tr>
<tr>
<td>ABCA2</td>
<td>E35/T5562C</td>
<td>Exon 35</td>
<td>His/His</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>ABCB1</td>
<td>E21/G2677T/A</td>
<td>Exon 21</td>
<td>Ala/Ser/Thr</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>ABCB4</td>
<td>E9/C504T</td>
<td>Exon 9</td>
<td>Asn/Asn</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>ABCB10</td>
<td>I1/T1541G</td>
<td>Intron 1</td>
<td>T</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCB11</td>
<td>E24/A3084G&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Exon 24</td>
<td>Ala &gt; Ala</td>
<td>G</td>
<td>-</td>
</tr>
<tr>
<td>ABCC1</td>
<td>5'FR/G-260C</td>
<td>5'FR</td>
<td>G</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCC2</td>
<td>E1/C-24T</td>
<td>Exon 1</td>
<td>T</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCC4</td>
<td>I4/C-3695G</td>
<td>Intron 1</td>
<td>T</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCC6</td>
<td>I3/A-1022C</td>
<td>Intron 30</td>
<td>G</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCC7</td>
<td>I2/G7A</td>
<td>Intron 23</td>
<td>T</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCC8</td>
<td>I1/A953C</td>
<td>Intron 1</td>
<td>T</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCC9</td>
<td>E21/A3986T</td>
<td>Exon 21</td>
<td>Ser/Thr/Thr</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>ABCC10</td>
<td>I1/A953C</td>
<td>Intron 1</td>
<td>T</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABCD2</td>
<td>E5/C421A</td>
<td>Exon 5</td>
<td>Glu/Lys</td>
<td>T</td>
<td>-</td>
</tr>
</tbody>
</table>

<sup>a</sup>LRH represents SNPs that passed the modified long-range-haplotype test only if not corrected for multiple testing; MTC represents SNPs that passed both the modified LRH test and subsequent multiple test correction.

<sup>b</sup>iHS values of SNPs were obtained from Haplotter. iHS values in bold indicate SNPs that also passed the LRH test for RPS in this study. ASN indicate aLRH represents SNPs that passed the modified long-range-haplotype test only if not corrected for multiple testing; MTC represents SNPs that passed both the modified LRH test and subsequent multiple test correction.

The distribution of RPS SNPs occurred earlier in the CHB and JPT populations compared to the CEU and YRI populations, resulting in the fixation of these advantageous alleles in the CHB and JPT populations. **ABCC11** was also found to be a high F<sub>ST</sub> candidate selection gene (see earlier paragraph and Table 1).

Except for the JPT, the other three populations had between 20% and 50% (32/64) RPS SNPs residing in five (YRI) to nine (CEU) ABC genes. Curiously, there was only one RPS SNP, c9/C504T in **ABCB4**, in the JPT population. The reason for this observation is unknown, although it could be an artifact arising from a bias in the selection of SNPs used in the HapMap project.

Most of the ABC genes contained not more than three SNPs that exhibited evidence of RPS, except for **ABCC2** and **ABCC11** which contained seven and four RPS SNPs, respectively (Table 2). Nine of the 32 RPS SNPs reside in the protein-coding region of the genes, although only four of these nine SNPs result in a non-synonymous amino acid change (Table 2). Three other RPS SNPs reside in potential regulatory regions such as the promoter (5'FR/G-260C in **ABCC1**; c1/C-24T in **ABCC2** and 5'FR/A-59G in **ABCC13**), while another two occur within intron/exon boundaries (i17/T20G in **ABCC5** and i1/G7A in **ABCC10**). Hence, ~44% (14/32) of the RPS SNPs are sited within potentially functional regulatory regions of genes. The remaining 56% of RPS SNPs reside in introns, which have less obvious functional significance. These SNPs may reside within undetected intron splice enhancers, thus affecting splicing efficiency, or they may be in LD with unobserved causal SNPs.

When Type I error reduction was performed, only 13 SNPs residing in 10 ABC genes (**ABCA2**, **ABCB4**, **ABCC2**, **ABCC3**, **ABCC7**, **ABCC8**, **ABCC9**, **ABCC10**, **ABCC11** and **ABCG2**) remained positively selected (Table 2). Only ~40% of the identified RPS SNPs remained positively selected after Type I error reduction. The distribution of RPS SNPs in the different populations was also found to be different after Type I error reduction. In the CEU, only ~14.3 and 22.2% of the previous number of RPS SNPs and genes, respectively, remained positively selected after undergoing multiple test correction. In the CHB, ~42 and 56% of the SNPs and genes, respectively, showed continued evidence of RPS after multiple test correction. On the other hand, 60 and 80% of the SNPs and genes, respectively, in the YRI were still positively selected after multiple test correction.
Figure 1. Schematic representation of the 18 ABC transporter genes and positions of analyzed SNPs. Each gene is represented by a thick horizontal line, and vertical bars represent exons. Gene, exon and intron sizes are drawn to scale. Empty boxes at gene ends represent 5' and 3' untranslated regions. Rs numbers indicate dbSNP IDs. SNPs that exhibit RPS even after correction for multiple testing are boxed by solid lines. SNPs that exhibit RPS only when multiple testing is not taken into consideration are boxed by dotted lines. Overhead lines connecting different SNPs indicate SNPs in tight LD with each other ($r^2 > 0.8$).
This is consistent with earlier observations that the YRI showed strongest signals of positive selection by several measures (17). In marked contrast, the only positively selected SNP in the JPT was no longer positively selected after correction for multiple testing.

Of the 13 positively selected SNPs remaining after MTC, one resides at an exon/intron boundary (SNP i1/G7A in ABCC10) while another four reside in the protein coding region, of which two result in non-synonymous amino acid changes. Similar to the results obtained prior to multiple test correction, only ~40% (5/13) of RPS SNPs reside on potentially functional regions of the gene.

**Experimental validation of functionality of two RPS SNPs**

Functionality of two RPS SNPs was experimentally assessed. One of these SNPs, e1/C-24T in ABCC2, showed evidence of RPS only in the absence of MTC while the other, i27/C-1020G in ABCC9, is positively selected even when MTC was performed (Table 2). As SNP e1/C-24T in ABCC2 resides at the promoter region, we performed promoter-reporter assays in four different cell-lines to evaluate if the C- and T-alleles of this SNP confer differential promoter activity. As shown in Figure 2A, the T-allele promoter demonstrated higher activity compared to the C-allele promoter in all cell-lines.
The most significant difference was observed in the HeLa (cervical adenocarcinoma) and HCT116 (colon carcinoma) cell-lines (P  <  0.001), followed by the Hep3B (hepato-cellular carcinoma) cell-line (P  <  0.01), and finally the embryonic kidney HEK-293 cell-line (P  <  0.05). We next evaluated the correlation between this SNP e2/C-24T genotype and ABCC2 transcript levels in human liver samples. As only one liver sample with a homozygous TT genotype was available, the result for that sample was combined with those from heterozygous CT samples. As shown in Figure 2B and D, liver samples with a CT/TT genotype showed significantly higher ABCC2 mRNA expression when compared with liver samples with the CC genotype (P  <  0.05). From these experimental results, we conclude that this promoter SNP, which yield evidence of RPS only if multiple test correction was not performed, does indeed have a functional effect, at least in some tissues.

The second SNP that we functionally validated is the ABCC9 intronic SNP i27/C-1020G, which yielded evidence of RPS after multiple test correction (Table 2). We hypothesized that although it is an intronic SNP, it may result in differential ABCC9 transcript expression by affecting transcript stability, etc. We thus examined the correlation between ABCC9 genotype and ABCC9 transcript levels in human liver samples. As evident in Figure 2C and D, the positively selected G-allele was found to be correlated with a significantly higher ABCC9 transcript level expression (P  <  0.05), again suggesting that RPS SNPs identified by the LRH method have a functional effect.

DISCUSSION

SNPs at the 18 ABC transporter gene loci showed a diversity of allele frequencies, LDs and haplotypes amongst the four different populations. Generally, the YRI population displayed the weakest LD and greatest haplotype diversity. Some of the ABC genes were found to display strong LD and low haplotype diversity while others had weaker LD and higher haplotype diversity.

The LRH algorithm was employed to identify potentially functional SNPs displaying genomic evidence of RPS. When Type I error reduction was not performed, at least one SNP in each ABC gene was found to show evidence of RPS resulting in a total of 32 RPS SNPs. Except for the JPT populations, the other populations has 10–14 RPS SNPs residing in five to nine genes. Only one RPS SNP at the ABCB4 locus was identified in the JPT population. Approximately Forty-four percent of the RPS SNPs are sited within potentially functional regions of genes including exons, regulatory regions or intron/exon boundaries. When MTC was performed to reduce Type I error, only 13 SNPs residing in 10 ABC genes remained positively selected.
Comparison of results from this study with recent large-scale, genome-wide and gene-specific positive selection studies

Recently, several large-scale genome-wide (16–18) and gene-specific studies (19,20) were reported that also aimed to identify genomic regions exhibiting ‘signatures of positive selection’. Three of these studies (17–19) employed the LRH test or modifications thereof to examine ‘signatures of positive selection’. In one study, 81 candidate gene loci for skin pigmentation were examined using phylogenetic and FST approaches, and only a few positively selected loci were identified (20). Another study utilized four different tests including MAF, derived allele frequency (DAF), FST versus heterozygosity and LRH to study 168 immune-related gene loci, and six genes showed some evidence of positive selection (19). Of the six genes identified to be positively selected via the LRH test, only ABC1C1 and V413 can be validated using independent data from HapMap. Although ABC1B1 and ABCG2 were also included amongst the 168 genes examined, that study identified strong signals of RPS only at the ABC1C1 locus in the European population (19), which we similarly detected in the CEU before Type I error reduction (Table 2).

Wang et al. (16) employed the LD decay (LDD) approach to analyze 1.6 million Perlegen database SNPs spanning the entire human genome and reported that ~1.6% of SNPs residing in ~1800 genes may be under positive selection. Zhang et al. (18) employed a combination of another modification of the LRH test known as whole genome LRH (WGLRH) and flipped SNPs determination to interrogation 116 204 SNPs in the Asian, Caucasian and African American populations genotyped using the GeneChap Mapping 100 K high-density SNP array (Affymetrix, CA, USA). They identified several genes in each population that may be under RPS. One of these positively selected genes in Caucasians is ABCC4 [Supplementary Material, Table S3 of Zhang et al. 2006 (18)], which we similarly detected in the CEU using the new HapMap genotype data release, again before Type I error reduction (Table 2).

Voight et al. (17) utilized another modification of the LRH test known as integrated Haplotype Scores (iHS) to identify variants under RPS, using Phase I HapMap genotype data (release 16C1, June 2005) of ~800 000 SNPs in 309 unrelated individuals. They identified ~250 regions that showed evidence of recent selection in each population and developed a set of tSNPs for these regions (17). As both the Voight et al. and our current studies utilized data from HapMap and employed variations of the LRH test to search for signatures of positive selection, a comparison of results from both the studies was made (Table 2). In the Voight study, the CHB and JPT populations were combined into an ASN population, resulting in only three (ASN, CEU and YRI) rather than four populations. Analogous to our study determining the eEHJ for SNPs, the Voight study calculated the iHS for SNPs. Among the SNPs analyzed by Voight et al., iHS values for at least one population were available for 28 of our 32 RPS SNPs (before Type I error reduction) (http://hg-wen.uchicago.edu/selection/haplotter.htm) (17) (reproduced in Table 2). The authors reported that the most extreme 1% of SNPs have |iHS| values >2.5. Only four of these 28 SNPs (e25/G3542T in ABCC2, i10/G377T in ABCC7 and i27/C-1020G and i37/C533G in ABC9) had |iHS| values >2.5 and hence were also considered positively selected in the Voight study. These four SNPs are among our 13 SNPs that remained positively selected after correction for multiple testing.

A possible reason why our other RPS SNPs, after Type I error reduction, were not identified to be positively selected in the Voight et al. study could be due to the differences in the computational analyses between the two studies. While our study calculated eEHJ around 200 kb from the core SNP allele, the Voight study computed the integral of the observed decay of EHH from the core SNP allele until the EHH value drops to 0.05. Although the corresponding |iHS| values of our other RPS SNPs were less than 2.5, the trend of |iHS| values across populations was similar. With the exception of five SNPs (e24/A3084G in ABCB11, e1/C-24T in ABCC2, i30/A-31G in ABCC6; i1/G7A in ABCC10 and 5’FR/A-59G in ABCC13), populations in which a SNP was found to be positively selected in our study (with/without correction for multiple testing) also had higher |iHS| values when compared with populations in which the SNP was not selected for.

Evidence of functionality of RPS SNPs from the published literature

Two ABC genes ABCC1 and ABCC4 that we found to exhibit RPS only when MTC was not performed, were also identified as strongly selected for in large-scale gene-specific (19) or genome-wide (18) studies, suggesting perhaps that in our effort to correct for possible Type I error, we may have inadvertently increased Type II error rates and discarded potentially functionally important SNPs. As genetic evidence for RPS would be greatly strengthened by functional evidence (15), we employed two approaches to evaluate the functional significance of some of the ABC transporter SNPs that we identified to be under RPS. Firstly, we performed a survey of previous publications reporting an association between any of these RPS SNPs and any function, drug response or disease. Several publications have examined the association of SNPs within various ABC transporter genes and function, drug response as well as disease susceptibility.

A synonymous SNP, e15/C2127T (rs908832), in exon 15 of the ABCA2 gene was found to be associated with early onset Alzheimer’s disease (AD) (28–30). However, this SNP is not genotyped even in the Release 21 of HapMap (June 2006). The most extensively studied ABC gene is the ABCB1 or MDR1 gene. SNPs e21/G2677T/A and e26/C3435T in MDR1 were found to be in high LD (25,31). In a SNP genotype analysis across 200 kb of the MDR1 gene in five populations, we previously reported statistically significant evidence of RPS, without correcting for multiple testing, for the e21/G2677T and e26/C3435T alleles in the Chinese population, and for the e26/C3435T allele in the Malay population (25). In the present study using HapMap genotype data, SNP allele e21/G2677T was again found to be under RPS in the CHB, and also CEU, populations when not corrected for multiple testing (Table 2). No genotype data for SNP e26/C3435T was available even in Release 20 of the HapMap project. Nonetheless, although both MDR1 SNPs only exhibited evidence of
RPS when there was no Type I error reduction; the functional significance of these two SNPs is evident from various reports. 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) and/or e1/-129(T/C), with differences in MDR1 expression and function, drug response and clinical outcome (reviewed in 32–34), and with susceptibility to various diseases (35–43).

The T-allele SNP e9/C504T in ABCB4, which showed evidence of RPS in CEU when there was no Type I error reduction, was reported to be significantly lower ($P < 0.001$) in frequency in low phospholipid-associated cholelithiasis (LPAC) syndrome patients who do not carry disease causing mutations (DCM), when compared with the control population (44). Several ABCB4 and ABCB11 polymorphisms which occur at low frequency ($<10\%$) in control populations, as well as DCM not found in healthy controls, have also been reported to be associated with various phenotypes (45–47). However, our RPS-detection strategy to identify functionally important polymorphisms focuses on high frequency SNPs and will not identify functionally significant but low frequency SNPs or DCM.

Although the RPS SNP e24/A3084G in ABCB11 is not known to be associated with any functional difference, SNP 3’-UTR/A236G, which is in strong LD ($r^2 = 0.97$) with e24/A3084G in the CEU, has been reported to be associated with intrahepatic cholestasis of pregnancy (48).

A few studies have also examined some polymorphisms within the MRP1/ABCC1 gene and have thus far found no association between the few ABCC1 SNPs tested and any functional differences (49–52). Nonetheless, two very low frequency ($<2\%$) non-synonymous ABCC1 SNPs, e22/G2965A (Ala989Thr) and e10/G1299GT (Arg433Ser), were reported experimentally to result in decreased estradiol 17B-glucoronide (53) and organic anion transport but increased doxorubicin resistance (54), respectively. These functionally significant but very low frequency SNPs would not have been identified with our high-SNP-frequency RPS strategy. However, none of these studies examined SNP 5’FR/G-260C, which we found to exhibit RPS when Type I error reduction was not performed (Table 2). Since the ABCC1 locus has weak LD and high haplotype diversity (Table 1) (27), this may account for the failure of the previous studies to detect association between other ABCC1 polymorphisms and functional differences. We recently genotyped several ABCC1 SNPs in five different populations and also identified 5’FR/G-260C as positively selected, even after correction for multiple testing. We further experimentally demonstrated allelic differences in MRP1 promoter activity in four different cell-lines (27). A possible reason why, in this current study, 5’FR/G-260C failed the positive selection test after correction for multiple testing could be due to the increased density of ABCC1 SNPs examined in the present study, which may have increased the haplotype diversity, thus reducing the EHH and rEHH values of 5’FR/G-260C. Nonetheless, experimental demonstration of the functionality of this SNP suggests that under certain conditions, Type II error may increase when multiple test correction to reduce Type I error is performed.

Two ABCC2 SNPs (e1/C-24T and e10/G1249A) that exhibited RPS when Type I error reduction was not performed (Table 2) have been previously examined in some studies. The promoter SNP e1/C-24T was reported not to be a risk factor for the development of spina bifida (55). This SNP was also found not to affect the ABC22 expression in human placenta (56) or duodenal enterocytes (52). However, e1/C-24T had an effect on ABC22 mRNA expression in normal kidney cortex (57). The exonic SNP e10/G1249A, which involves a conservative amino acid change (V471I), was reported to result in decreased ABC22 mRNA expression in preterm babies (56). In addition to these two SNPs, a few other ABC22 SNPs were also reported to be associated with functional differences (58). However, these SNPs were either not genotyped in the HapMap Project or exhibit very low MAF (~1\%).

The ABCC3 promoter SNP 5’FR/C-211T (rs4793665), which was not examined in this study as it was not genotyped in HapMap (Release 21, Jun 2006), was reported to alter hepatic mRNA expression (59), but not ABCC3 promoter activity (60) or mRNA expression in acute leukemia patients (61).

The ABCC6 SNP 130/A-31G, which exhibited RPS when Type I error reduction was not performed is in relatively high LD with the non-synonymous SNP e27/G3826T in the four populations ($r^2 = 0.37–0.66$). The latter SNP has been reported to be associated with plasma lipoprotein variation in Canadian Inuit (62). Another ABCC6 SNP 5’FR/A-219C, which was not genotyped in HapMap (Release 21, June 2006), was also found to be associated with Pseudoxanthoma elasticum infection (63).

Recently, the ABCC11 SNP e5/G538A was found to be a determinant of earwax type, with the G- and A-alleles corresponding to wet and dry earwax types (64). The A-allele was also found to result in a lower excretory activity for cGMP compared to the G-allele (64). Interestingly, based on the ‘Out-of-Africa’ hypothesis, the ancestral allele for this SNP should be the G-allele, as it is the only allele in the YRI population (Supplementary Material, Table S2). The A-allele probably arose in the CEU, occurring at a frequency of 12.71%, and rapidly rose to frequencies of >88% in the Asian populations, likely as a result of strong selective pressure. However, this is the only HapMap Project genotyped ABCC11 SNP which was polymorphic in the Asian population (Supplementary Material, Table S2). All other HapMap genotyped ABCC11SNPs are monomorphic in the Asian populations (Supplementary Material, Table S2), thus preventing the LRH test from being performed on SNP e5/G538A. Nonetheless, this SNP was found to have very high $F_{ST}$ of 0.81 (Supplementary Material, Table S2) and represents the highest $F_{ST}$ value out of 571 total SNPs at the ABC gene loci, suggesting that this gene may be under very strong selection pressure.

Polymorphisms within ABCG2 have also been quite extensively examined for functional significance. SNP e5/C421A, which we found to be under RPS even after multiple test corrected (Table 2), results in a non-synonymous amino acid change (Q141K) and resides in the functionally important ATP-binding region between the Walker A and B motifs. It has been suggested that this SNP may alter the ABCG2 protein structure, making it more susceptible to degradation (65). The positively selected A-allele of this SNP was found to result in markedly decreased protein expression (65–67) but not mRNA expression (65). This polymorphism was also found to alter ATPase activity (68,69) and drug efflux properties.
(68,69), increasing the bioavailability of drugs (70,71). These studies affirm that RPS SNP e5/C421A has functional significance.

In summary, for six of the 32 ABC SNPs that we detected evidence of RPS, there is good corroborating evidence in the literature to support a functional effect (Table 3). Of these six RPS SNPs, only one retained evidence of RPS after correcting for multiple testing, suggesting that by reducing the possibility of Type I error, the rate of Type II error may be raised to a counter-productive level.

Although a total of 82 single nucleotide sequence changes at these ABC transporter genes have been reported in the literature to be functionally significant, the majority of these sequence changes (80% or 66/82) are DCM, which are not observed in the normal control population or occur at low frequency (10%) in the control population. These functionally important sequence changes will not be detected using the current algorithm of RPS detection, which focuses on common sequence changes that occur at high frequency in the normal population. An additional 8.5% (7/82) of the functionally significant polymorphisms reported in the literature were not genotyped in HapMap (Release 21, June 2006), and hence were not identified in this study. Of the remaining nine functionally significant SNPs which occur at >10% frequency in the normal population and were also genotyped in HapMap, six were found to show evidence of RPS, with another two in tight LD with RPS SNPs. Hence nearly 90% (8/9) of the reported functionally significant SNPs that were genotyped in HapMap and occur at >10% in the normal population demonstrate evidence of RPS, or are in tight LD with SNPs that show RPS. The only sole HapMap genotyped, high frequency SNP that did not show evidence of RPS in this study is the ABCC11 SNP e5/G538A. As discussed earlier, this was because the LRH test cannot be performed on this SNP. However, this SNP is also very likely under strong selective pressure as it displayed the highest $F_{ST}$ value among 571 ABC SNPs tested (Supplementary Material, Table S2).

### Evidence of functionality of RPS SNPs from prospective experimental validation

In addition to published evidence of the functionality of SNPs identified as positively selected, we further tested two RPS SNPs for functionality by experimental means. The first RPS SNP we examined was ABCC2 promoter SNP e1/C-24T, which was found to be positively selected only if no MTC was performed (Table 2). As discussed in the previous section, a functional effect of this SNP was previously observed in normal kidney tissues (57) but not in human placenta (56) or duodenal enterocytes (52). Promoter-reporter assays revealed that the RPS T-allele of SNP e1/C-24T mediated higher ABCC2 promoter activity in several cell-lines tested (Fig. 2A). The higher promoter activity of the T-allele is corroborated by evidence that SNP e1/C-24T CT/TT genotypes correlated significantly with higher ABCC2 mRNA expression ($P < 0.05$) in human livers (Fig. 2B and D). Hence this promoter SNP, which showed evidence of RPS only if MTC was not performed, does indeed have a functional effect. We also validated the functionality of ABCC9 intronic SNP i27/C-1020G, which yielded evidence of RPS even after MTC (Table 2). The positively selected G-allele of this SNP mediated significantly higher ABCC9 mRNA levels ($P < 0.05$) in human livers (Table 2C and D), again demonstrating a functional effect of RPS SNPs identified by the LRH method.

---

**Table 3. Summary of functionally significant SNPs reported in the literature**

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of RPS SNPs (this study)</th>
<th>No. of functionally significant SNPs reported in literature</th>
<th>MAF &lt;10%, including DCM</th>
<th>MAF &gt;10% Evidence of RPS</th>
<th>Tight LD with RPS SNPs</th>
<th>No evidence of RPS</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCA2</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(28,29)</td>
</tr>
<tr>
<td>ABCB1</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(74)</td>
</tr>
<tr>
<td>ABCB4</td>
<td>3</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>3</td>
<td>–</td>
<td>(44,47,75)</td>
</tr>
<tr>
<td>ABCB11</td>
<td>1</td>
<td>–</td>
<td>3</td>
<td>–</td>
<td>1</td>
<td>–</td>
<td>(47,48,76)</td>
</tr>
<tr>
<td>ABCC1</td>
<td>1</td>
<td>–</td>
<td>2</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>(54,77)</td>
</tr>
<tr>
<td>ABCC2</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>(56–58,78)</td>
</tr>
<tr>
<td>ABCC3</td>
<td>2</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(59–61)</td>
</tr>
<tr>
<td>ABCC4</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCC5</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCC6</td>
<td>1</td>
<td>1</td>
<td>42</td>
<td>–</td>
<td>1</td>
<td>–</td>
<td>(62,63)</td>
</tr>
<tr>
<td>ABCC7</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>(79)</td>
</tr>
<tr>
<td>ABCC8</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCC9</td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCC10</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCC11</td>
<td>4</td>
<td>1</td>
<td>42</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>(64)</td>
</tr>
<tr>
<td>ABCC12</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCC13</td>
<td>2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ABCG2</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>1$^a$</td>
<td>–</td>
<td>–</td>
<td>(65,80)</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>7</td>
<td>66</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

RPS: recent positive selection; DCM, disease causing mutations.

$^a$ Showed evidence of RPS even after MTC, multiple testing correction.
In summary, through the detection of signatures of RPS, we have identified 32 SNPs in 18 ABC transporter genes with potentially functional importance. Some of these RPS SNPs have been previously reported to be functionally important, and we have further validated the functionality of two RPS SNPs. Although the functional importance of the remaining RPS SNPs remain to be experimentally confirmed, their presence within important drug/xenobiotic transporter genes nonetheless suggests likely involvement between these genes and drug response and/or association with complex disorders that are mediated by xenobiotic exposure. In contrast to using putative RPS SNPs for association testing, the tSNP method requires many more SNPs to be tested to minimize Type II errors, which in turn increases Type I error rates, unless correction for multiple SNP testing is performed.

Since the traffic of a single drug can be regulated by more than one ABC transporter, the simultaneous evaluation of these 32 SNPs from 18 ABC transporter genes may provide more powerful insights into the genomic basis for specific drug response variation or complex disease association than possible with current strategies. With the development of more complex algorithms to analyze the fractional contribution of each SNP variant to drug response, it becomes more attractive to assess a comprehensive set of functionally significant SNPs representing all genes impacting drug response (including drug transporters, drug metabolizing enzymes, transcriptional regulators such as SXR, drug receptors and others), so that a more complete understanding of the genomic determinants of variation in drug response can be achieved.

MATERIALS AND METHODS

Genotype data

Genotype data of SNPs within and around (200 kb upstream and 200 kb downstream) the 18 ABC drug transporter gene loci were extracted from the International HapMap Project Phase 1 (http://www.hapmap.org) (Release 16C.1, June 2005). A total of four populations were examined in this study including 60 unrelated parents of US residents with Northern and Western European ancestry from the Centre d’Etude du Polymorphism Humain (CEPH), 60 unrelated parents of the Yoruba people of Ibadan, Nigeria (YRI), 45 unrelated Japanese from the Tokyo area (JPT) and 45 unrelated Chinese from Beijing (CHB). SNPs with low MAF (<0.05) in all four populations were removed from further analyses.

Haplotype and linkage disequilibrium (LD) profiles

The haplotype and LD profiles at the 18 ABC drug transporter gene loci in the four populations were examined utilizing intragenic SNPs as well as SNPs within ~2 kb upstream of each gene, representing the promoter regions. In each population, SNPs with MAF greater than 5% were utilized to determine the LD between SNP pair. LD and half LD were determined as previously described (25). For the determination of the haplotype profile, SNPs with genotype data for all four populations and with MAF >5% in at least one population were selected to ensure that the number and identity of the SNPs for haplotype analyses were the same for all four populations. Haplotype was inferred using the SNPPhap program (http://www.gene.cimr.cam.ac.uk/clayton/software/), developed by David Clayton, which utilizes the Expectation-Maximization (EM) algorithm to estimate the frequencies of haplotypes from large numbers of biallelic SNPs from unphased genotype data of unrelated subjects.

Identification of signatures of positive selection using the long-range-haplotype test

At each of the 18 ABC drug transporter gene loci, SNPs within as well as 200 kb upstream and 200 kb downstream of each gene were selected for LRH analyses based on the following criteria. (1) For exons, exon-intron boundaries and the 2 kb upstream region of each gene, all polymorphic SNPs were included; (2) for intronic regions, LD was determined for all polymorphic SNPs. All SNPs that were not in strong LD with other SNPs ($r^2 < 0.8$) were included in the analyses. However, for SNPs in tight LD with each other ($r^2 \geq 0.8$), only one representative SNP was included. The representative SNP was selected based on the quality of the HapMap genotyped data for that SNP in the population; (3) for the 200 kb upstream and downstream regions, SNPs were selected at a density of approximately one SNP per 10–20 kb of sequence. Haplotype frequencies were then estimated from the selected SNPs within and around the gene using SNPPhap. A modified LRH test was then employed to evaluate the evidence of RPS based on haplotype frequencies (25). Extended haplotype homozygosity (EHH) and relative EHH (rEHH) were plotted against genomic distance and allele frequencies, respectively, as described previously (25). Coalescent simulations were performed on SNPs with high rEHH values to validate the evidence of positive selection. Four different population models namely, constant population size model, expansion model, extreme bottleneck model and highly structured population model were simulated as described previously (25). MTC to reduce Type I error is accomplished using a simulation approach to obtain the exact critical value for the LRH test as described previously (27).

Functional assessment of positively selected SNPs

We assessed two candidate positively selected SNPs, $ABCC2\ e1/C-24T$ within the promoter, and $ABCC9\ i27/C-1020G$ within intron 27. As both $ABCC2$ and $ABCC9$ are expressed at moderate levels in the liver (Supplementary Material, Table S1), we utilized human liver tissues to evaluate the association between different alleles of these SNPs and liver transcript levels.

Human liver tissues were obtained from the National Cancer Centre Tissue Repository with prior approval from the Institutional Review Board (IRB). Genomic DNA was extracted from human liver tissue using the Puregene DNA Isolation Kit (Gentra Systems, Minneapolis, MN, USA) according to manufacturer instructions. Purified DNA was dissolved in water, quantitated and stored at $-70^\circ$C. Total RNA from liver tissue samples was isolated using the TRI-Reagent Kit (Mregene, Cincinnati, OH, USA) as per manufacturer’s
protocol. cDNA was synthesized by reverse transcription of total RNA using Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA, USA) as per manufacturer’s protocol.

Genotyping of the ABCC2 and ABCC9 candidate positively selected SNPs on genomic DNA samples was accomplished by multiplex PCR amplification and multiplex minisequencing, as described previously (72).

As ABCC2 e1/C-24T is a promoter SNP, we also performed an in vitro expression analysis to determine if different alleles would result in differences in ABCC2 promoter activity. A 1259 bp wild-type promoter fragment of the ABCC2 gene (NM_000392.1) spanning from −1229 to +30 (+1 denotes the translation start site) was amplified and inserted upstream of the β-galactosidase (β-gal) reporter gene in an expression vector which also contains the enhanced green fluorescent protein reporter gene driven by the cytomegalovirus promoter for normalization against differences in transfection efficiency. The alternative T-allele of e1/C-24T was recapitulated via in-vitro PCR mutagenesis. Plasmid constructs were sequenced across the PCR amplified regions to exclude PCR-induced nucleotide mis-incorporations prior to use. Transfection of constructs, as well as the promoter reporter assay, was performed as described previously (27,73).

ACKNOWLEDGEMENTS

This work is supported by grants from the National Medical Research Council (NMRC) of Singapore: Academic Research Council, Ministry of Education (ARC, MOE) of Singapore and Academic Research Fund (ARF) of the National University of Singapore to Caroline G. Lee as well as the NMRC block grant through the National Cancer Centre to Caroline G. Lee.

Conflict of Interest statement. None declared.

REFERENCES


