I GENERAL INTRODUCTION

Chapter 1 Background

1.1 Overall objectives of thesis

The successful completion of the human genome project has created an enormous amount of information and also generated many scientific insights about future drug therapy. One major interest that has risen over the years is the possibility of personalized medicine. Pharmacogenomics is the driving force of personalized medicine that uses genetic analysis to understand the interaction between drug therapy and the genetic makeup of an individual (Evans and McLeod 2003). This information may allow us to design individual-based medicine to increase efficacy, reduce side effects and avoid adverse drug reactions (ADR).

One of the most significant works to date in the area of pharmacogenomics is the discovery of Single Nucleotide Polymorphisms (SNPs). There is limited sequence variation between any two individuals so there is great interest in identifying these unique genetic differences. Of these genetic differences, over 90% are single nucleotide polymorphisms (Deloukas and Bentley 2004). A SNP is a single base substitution of one nucleotide with another, and is observed in the population at a frequency of at least 1%. There are concerted efforts to genotype all human variations, especially SNPs across the human genome in large populations. By studying SNP profiles, genes associated with any human trait such as disease susceptibility or aberrant drug response can be revealed. Association studies can detect and indicate which particular SNP profile is most likely to be associated with the genes of interest. Eventually, panels of SNP profiles that are characteristic of a variety of diseases will be established and therefore be used for screening individuals for susceptibility to disease or drug sensitivity by analyzing their DNA. It is envisaged that

pharmacogenomics has the potential to predict interindividual pharmacokinetic differences by genetic polymorphisms of drug transporters/enzymes or pharmacokinetic changes by transporter-mediated drug interactions (Evans and McLeod 2003). By predicting the bioavailability profiles based on this genomic information, drug toxicity such as adverse drug reactions (ADRs), and loss of drug efficacy due to underdosage, would be prevented.

While SNPs are abundant in the human genome, the majority of these SNPs may be neutral or benign variations. Therefore the next challenge is to search for the causal variants responsible for inter-individual variability in drug response or disease susceptibility. This can be facilitated by rudimentary knowledge of target genes and the discovery of novel technology for mapping genetic variation in individuals. Keeping in view that SNP profiles of drug transporters in individuals and populations are variable, this thesis has 2 main objectives:

1. To develop a cost-effective SNP genotyping method for the purpose of mid- to high-throughput analysis of multiple SNPs;
2. To characterize the haplotype and linkage disequilibrium profiles of 2 nucleotide analog transporters, *ABCC*4 and *ABCC*5, for future gene-based association studies;

As the objectives of this thesis cover a wide spectrum, at least one chapter is devoted to each objective in the sections on Methods and Materials, Results and Inferences. The final section of General Discussion closes this thesis with an overall discussion on results obtained and future perspectives.

1.2 SNP genotyping strategies

Single nucleotide polymorphisms are by far the most abundant form of genetic variation (about 90% of all human genetic variation), occurring every 100 to 300 bases along the 3-billion-base human genome (Cargill et al. 1999). Although more than 99% of human DNA sequences are the same across the population, it is thought that many complex human phenotypes have a significant genetic component and the variability is likely to be a result from differences in SNP genotypes (Meyer 2004). Given the importance of SNPs in pharmacogenetics and pharmacogenomics, it is no small wonder that there has been an impressive development in SNP genotyping technologies over the past few years, especially the push for more cost-effective high-throughput assays. DNA microarrays and mass spectroscopy are still too expensive for the average laboratory, so other alternative methods, preferably scalable, liquid-based and amenable for multiplexing have to be developed.

A genotyping strategy typically consists of 3 main components: target DNA amplification, followed by allelic discrimination and signal detection (Chen and Sullivan 2003). Most current methods are assortments of different methods of allelic discrimination and signal detection (Gray et al. 2000; Kwok 2001; Chen and Sullivan 2003).

Target DNA amplification is a critical step to generate enough copies of specific amplicons containing the SNP of interest. Most genotyping methods therefore begin with the Polymerase Chain Reaction (PCR). The Invader® technology marketed by Third Wave Technologies (Hsu et al. 2001) is currently the only method which is able to eliminate the need for amplification of sample DNA by genotyping directly from the human genome. Therefore the total cost of PCR can be of significance to most genotyping methods (Chen and Sullivan 2003). One way to increase efficiency is to

amplify several amplicons, each containing SNPs of interest, in a single reaction (Henegariu et al. 1997). Multiplex PCR is not without its problems. With multiple sets of primers, there may be nonspecific or uneven amplifications (Henegariu et al. 1997). The presence of gene families such as the ATP-Binding Cassette (ABC) Superfamily of transporters, as well as pseudogenes and other conserved sequences in the genome can give rise to false genotypes.

In some genotyping methods such as those that involve primer extension, a clean-up step needs to be performed to free the products of the PCR reaction from excess dNTPs and PCR primers. Generally, there are three ways to do a post-PCR cleanup for genotyping purposes. Extracting desired products via electrophoresis on an agarose gel is not considered a viable method here as quantity loss from small volumes of PCR reaction is not acceptable and this would also run contrary to the benefits of multiplex PCR. The first method of PCR clean-up is the use of magnetic beads for DNA binding and elution of PCR products after washing away primers. The second method uses ultrafiltration to retain PCR products via a spin column while primers and excess dNTPs are removed. The PCR products can be eluted off the membrane after washing. The last method, is to enzymatically cleave excess dNTPs and PCR primers using shrimp alkaline phosphatase (SAP) and E.coli exonuclease I (Exo I) respectively.

Allelic discrimination can be divided into 2 categories: sequence non-specific or sequence specific (Kwok 2001). Sequence non-specific methods of allelic discrimination such as heteroduplex analysis and Denaturing High-Performance Liquid Chromatography (dHPLC), make use of the different electrophoretic ability or molecular sizes of either heteroduplexes or single-stranded DNA molecules. Unless modified, they are not useful in genotyping known variants.

The assay chemistry of sequence-specific allelic discrimination methods includes allele specific hybridization, polymerase extension, oligonucleotide ligation, and enzymatic cleavage (Syvanen 2001). A different enzyme is used in each of these methods, namely DNA polymerases, DNA ligases, and structure-specific enzymes. Allele specific hybridization makes use of two allele-specific probes, each designed to anneal to the target sequence only if the both sequence and probe are perfectly complementary. In allele-specific extension (ASE) and allele-specific PCR (AS-PCR), DNA polymerases extend far more efficiently in matched sequences than unmatched ones. The allele-specific primers are differentially labeled with tags for determination of genotype. In Single-Base Extension (SBE) or minisequencing, DNA polymerase extends one base pair 3' of a probe designed to anneal immediately upstream of a polymorphic site. Differentially labeled ddNTP are used instead of dNTP to label and terminate the extension. In pyrosequencing, detection is based on the formation of pyrophosphate as a byproduct of DNA polymerization. DNA polymerase catalyzes the incorporation of the deoxynucleotide triphosphate into the DNA strand, if it is complementary to the base in the template strand. Each incorporation event is accompanied by release of pyrophosphate (PPi) in a quantity equimolar to the amount of incorporated nucleotide. The addition of dNTPs is performed separately so that signals can be correlated to calling of genotype. The ability of DNA ligases to repair minor nicks in DNA is made used of in Oligonucleotide Ligation Assays (OLA). Ligation of two adjacent oligonucleotides annealed on target DNA only occurs in the presence of DNA ligase and only if the oligonucleotides perfectly match the template at either end of the ligation sites. Lastly, enzymatic cleavage by specific enzymes can also be used for allelic discrimination. The 5' nuclease activity of DNA polymerases can cleave a probe annealing at a SNP

site. If the probe is not perfectly annealed to the template, it will not be cleaved. In the Taqman assay, two dually labeled probes are used for allele discrimination. Some DNA polymerases with endonuclease activity can cleave complexes formed from the hybridization of overlapping oligonucleotide probes. The probes are designed so that the polymorphic site is at the point of overlap. When there is no overlap as in the case of a primer with a one-base mismatch, there is no cleavage. The Invader assay makes use of 2 allele-specific oligonucleotides such that one of the two would be cleaved and release a 5' arm. The released 5' arm can be used in a secondary reaction for greater signal amplification (Hsu et al. 2001).

While a definite genotype of a known SNP variant can be ascertained with any of the methods, each of them has its own advantages and disadvantages. Various detection methods such as fluorescence, colorimetry, chemiluminiscence and mass spectrometry also influence choice of a genotyping platform (Kwok 2001; Chen and Sullivan 2003). All methods can be fairly robust. SBE or minisequencing is almost as specific as DNA sequencing because the chemistry is the same in both cases except that the products of SBE are only one base longer than the probes used. This means that both SBE and sequencing may be performed on the same automated platform. This can be extremely cost saving since an existing sequencing machine in a facility can also be used for genotyping SNPs as well as discovering new ones. SBE also requires the least number of probes and all probes used are normal or HPLC purified oligos that are by far cheaper than labeled probes. Pyrosequencing is quantitative and so may be used for pooling samples when analyzing genotypes in large populations (Sham et al. 2002). Furthermore, it is able to perform single-tube haplotyping and thereby genotype multiple close SNPs (Pati et al. 2004). OLAs are potentially highly specific but the slow rate of reaction and the large numbers of modified probes

required are two of its disadvantages. The Invader assay is able to genotype an SNP directly from genomic DNA without prior amplification. On the other hand, the Invader assay requires more DNA than most genotyping methods and the design and purity of the probes is crucial (Kwok 2001). Pati et al., 2004, have compared SBE, pyrosequencing and the Invader assay and found them all favorable for the use in high throughput genotyping with minor differences in accuracy, cost and throughput (Pati et al. 2004). Lee et al., compared 4 methods of SNP genotyping (ASE, SBE, OLA, and direct hybridization) objectively by using the same flow cytometer as platform. They found that SBE is the most robust assay but is also comparatively more expensive, due to its inability to genotype both alleles in a single reaction (Lee et al. 2004b). The ASE assay was both cheaper and simpler than SBE.

The choice of a genotyping strategy is affected by many considerations, amongst which are accurate results, cost per SNP/sample and ease of optimization/use. In this thesis, SBE or minisequencing was adopted as the genotyping strategy based on previous reports for its high accuracy (Syvanen 2001). Its ability to genotype multiple SNPs in a single reaction, scalability to 96 samples, robustness and cost-effectiveness were deemed critical for haplotype and linkage disequilibrium studies of moderate sample sizes. There is no need for the purchase of a dedicated machine for genotyping as the same platform was also used for dideoxy sequencing (still the 'gold standard' for verifying sequences). The laboratory was therefore able to perform sequencing for novel SNP discovery as well as SBE for large-scale genotyping of these SNPs using a single automated capillary electrophoretic platform.

1.3 Population-based genetic association studies

Complex genetic diseases are familial disorders that are not attributable to a single dominant or recessive gene but to several genes contributing to the disease in a multiplicative or additive fashion. Polygenic inheritance also plays a significant role in the display of inter-individual variability in the therapeutic response to current clinical medicine. The genes involved in complex genetic diseases and variable drug responses are therefore harder to detect due to the small effects of each of the multiple component genes involved. There has been a shift in recent years to meet these challenges from linkage analysis to population-based genetic association studies. (Reich et al. 2001; Goldstein et al. 2003). While linkage analysis studies have been successful for rare single-gene defects, they have not been equally adept at locating causal variants responsible for complex gene diseases.

The approach to designing a population-based genetic association study is remarkably simple. Allele frequencies of a genetic variation (e.g. Single Nucleotide Polymorphism or Haplotype) in a candidate gene are compared between 2 groups of individuals, affected cases and unaffected controls. By testing candidate variations, it is theoretically possible to identify all the variants responsible for complex genetic diseases or variable drug responses. However, it would be both time-consuming and expensive to make an exhaustive and comprehensive comparison of all SNP sites, even with today's technology and corroborative electronic databases. One of the advantages of using population-based genetic association studies is that individuals can be selected to match the study e.g. for age and gender. Risks of contribution from the environment or non-genetic background can be assessed (Goldstein 2001; Goldstein and Weale 2001). Small genotypic effects from a relatively modest collection of cases and controls can be detected. By and large, unrelated individuals

are easier to obtain in numbers compared to families, trios (father-mother-child), siblings or twins.

1.4 SNPs and Haplotypes

Haplotypes are defined as specific combinations of alleles on an individual chromosome (Hoehe 2003). Traditionally, SNPs within and around genes were tested to correlate candidate genes with disease. Doubts concerning single SNP approaches have recently surfaced with studies demonstrating better correlation of complex phenotypes with haplotypes rather than with single SNP variants (Davidson 2000; Drysdale et al. 2000; Hoehe 2003). The argument for the use of haplotypes over SNPs is based on the evidence that single SNP-based candidate gene studies may be statistically weak in complex phenotypes. True associations may not be reflected due to low significant power and negative correlation with a single SNP does not exclude a positive correlation with the gene of interest.

1.5 Haplotypes and linkage disequilibrium

Linkage disequilibrium occurs when the observed frequencies of haplotypes in a population deviate from the haplotype frequencies predicted by multiplying together the frequency of individual genetic markers in each haplotype (Zondervan and Cardon 2004) (Figure 1). When there is no such deviation, then the population is said to be in linkage equilibrium. This observation that there is non-random association of alleles at different loci is of importance. The basis of using LD mapping relies on the fact that haplotypes are often associated with reasonably common diseases that have complex genetic origins. Identifying haplotypes may therefore make it easier to link them to specific complex diseases.

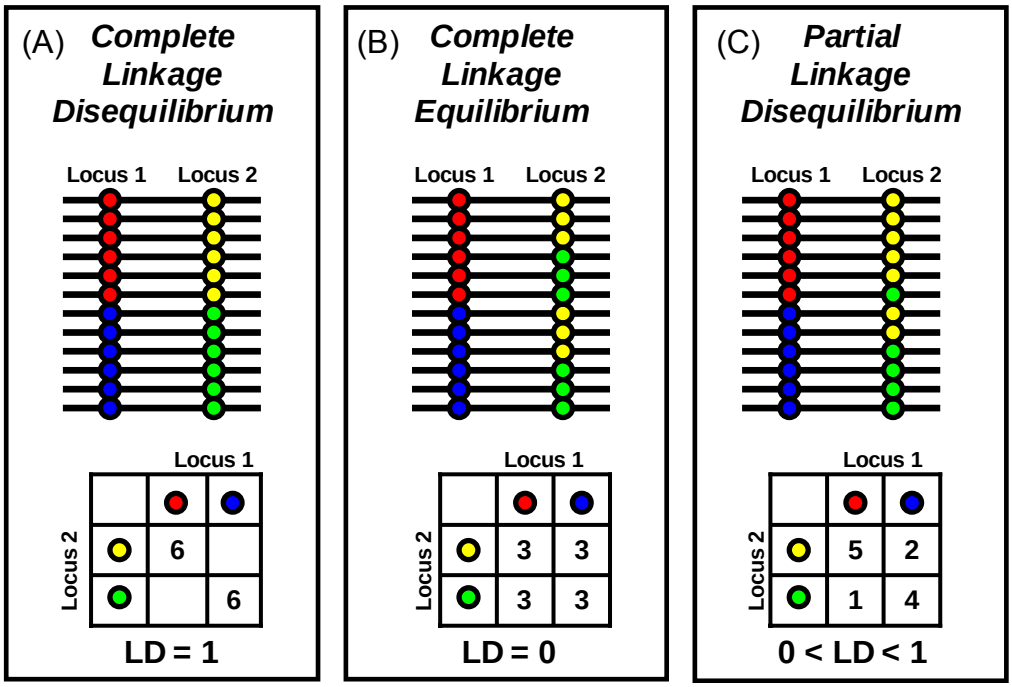


Figure 1. Linkage Disequilibrium and Linkage Equilibrium. Loci 1 and 2 are marker SNPs in close proximity. Panel A: When one allele (e.g. red) of Locus 1 is completely linked to another allele (e.g. yellow) of Locus 2, they are in complete linkage disequilibrium (i.e. LD = 1). Panel B: When the alleles of the two loci are completely unlinked, then complete equilibrium has occurred. Panel C: Alleles of the two loci are linked by different degrees, giving rise to a measurable LD value between 0 and 1.

The extent of LD is dependent on several factors both on the molecular level such as recombination and mutation rate as well as demographic and evolutionary factors such as migration, population growth and admixture between populations.

There are two main measures of LD, $|D'|$ and r^2. Both are based on the Lewontin's D. The first measure is derived from D, which measures the difference between the observed haplotype frequency and the expected haplotype frequency if the alleles are randomly segregating. *D'* is obtained by dividing D over Dmax, the maximum D possible for a given set of allelic frequencies at any two loci. When $|D'|$ is null, there is linkage equilibrium. When $|D'|$ equals to one, there is complete linkage disequilibrium (Zondervan and Cardon 2004). The main disadvantage of using $|D'|$ is that it can get highly inflated, especially in small sample sizes. It is also sensitive to

allele frequencies and can be inflated in SNPs with rare allele frequencies (Zondervan and Cardon 2004). The 'half length' LD is the distance at which the average *D'* value drops below 0.5 (Zondervan and Cardon 2004).

The next measure of LD, r^2, determines the correlation of alleles at two loci. When r^2 =0 there is perfect linkage equilibrium. When r^2=1 there is perfect linkage disequilibrium. r^2 has the advantage over |*D'*| in that it does not get inflated in small sample sizes and/or when the loci have rare alleles (Shifman et al. 2003). r^2 is therefore more appropriate when association studies are of interest. For pairs of biallelic markers, *D*' will be equal to 1.0 when one or two out of the four possible haplotypes are missing from the population, while *r* equals to 1.0 when there are only two haplotypes (Zondervan and Cardon 2004). When recombination does not occur between two markers, *D*' will be 1.0 (in the absence of mutation or genotyping error), while *r* will be dependent on allele frequencies of both markers. *D*' is therefore used as a model for recombination rates and *r* or r^2 as a model for association power.

The degree of LD between 2 alleles is dependent on how old the two polymorphisms are i.e. when they appeared in the populations and the degree of recombination between them. Reich et al., estimated the *D'* in the genome to be 60 kb by studying unrelated individuals in a United States population of north-European descent (Reich et al. 2001). The authors measured LD between two SNPs using the classical statistic *D*' as 'half length' LD. A study by Gabriel et al., estimated that half the human genome exists in blocks of 22 kb or larger in African and African American populations and 44kb or larger in European and Asian populations (Gabriel et al. 2002).

Markers that are in close proximity generally have stronger LD than those located far apart. Whilst there is some correlation of LD with genomic distance, it has been recently shown that there is also variation in the extent of LD. Demographic processes

can also affect LD. Population expansion can decrease LD (Kruglyak 1999) while population bottlenecks and population structure tend to increase LD.

1.6 Gene-based versus genome-based approaches

One of the famous Mendelian laws of inheritance states that hereditary factors are inherited independently (Independent Assortment). Therefore the presence of having one type of gene should not be influenced by the presence of another gene. In reality genes are often linked and thereby inherited together as a unit. As such, individuals often inherit stretches of DNA from their parents. The genome can therefore be segregated into segments of chromosomal sequence within which sets of common SNPs exist in high LD. Apparent sites of recombination called "recombination hotspots" punctuate these segments or "blocks" (Zhang and Jin 2003). This suggests that there is a haplotype block structure in the human genome. These haplotype blocks can thus serve as organization units to establish genome-wide LD mapping.

The length of each haplotype is of interest as it reflects the extent of LD. LD has been found to extend to 60-100 kbs in European populations as compared to a few kbs in African populations (Abecasis et al. 2001; Daly et al. 2001; Goldstein 2001; Patil et al. 2001; Reich et al. 2001; Rioux et al. 2001; Bonnen et al. 2002; Dawson et al. 2002; Gabriel et al. 2002). Within these haplotype blocks, there are a limited number of sets of observed haplotypes. Because of the limitation of genetic diversity withi each block, a small subset of common SNPs might be sufficient to define the majority of the haplotypes in any population.

1.7 Gene-based haplotypes

Just as genome-based haplotypes are based on combinations of alleles at sites across a chromosomal segment, gene-based haplotypes represent combinations of alleles at markers in a gene. These markers may be SNPs, Restriction Fragment Length Polymorphisms (RFLPs), Short-Tandem-Repeats (STRs) or microsatellites. Regardless of allele frequency, any causative allele in strong LD with any of the selected markers will be represented in the haplotypes. Gene-based haplotypes are much more precise than single SNP markers as they represent more heterozygosity and contain the entire LD structural information. As such, they will have greater power in elucidating the unobserved causative allele (Judson and Stephens 2001).

1.8 Haplotype reconstruction

To unambiguously determine haplotypes, one could either directly genotype pedigrees or use molecular methods in conjunction with genotyping individual samples if pedigrees are not available. Directly genotyping pedigrees relies on the fact that closely-located markers (as haplotypes) will be inherited as a unit unless separated by recombination events. Assigning haplotypes to pedigrees is laborious and expensive as well as increasingly difficult as the number of loci increases. Molecular and experimental methods are unambiguous methods for constructing haplotypes if pedigrees are not available. These methods include the allele-specific polymerase chain reaction (AS-PCR) and the use of somatic cell hybrids (Zhang et al. 2005). These techniques are used in small populations, as they are also laborious and expensive to perform on large number of loci or large sample pools.

Statistical inference of haplotypes from genotype data in large population studies thus remains the most viable method. To date, there are three principal computational

methods that haplotype inference software can be categorized into: parsimony, maximum-likelihood, and Bayesian (Lin et al. 2002; Crawford and Nickerson 2005). The Clark statistical algorithm utilized in the software programs HAPINFERX and HAPINF is based on parsimony. Clark's method (Clark 1990) first lists known haplotypes from unambiguous genotypes i.e. haplotypes in homozygous individuals and individuals with no more than one heterozygous locus. The algorithm then attempts to resolve ambiguous genotypes into one of the known haplotypes until they are inconsistent with any inferred haplotypes. It is not popular even though it is relatively fast and does not assume Hardy-Weinberg Equilibrium (HWE), because the level of ambiguity increases with the number of polymorphic sites considered or at which an individual is heterozygous (Niu 2004). The maximum likelihood is implemented via the expectation-maximization (EM) algorithm (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995) in the software programs Arlequin [http://lgb.unige.ch/arlequin/], HAPLO [http://krunch.med.yale.edu/haplo], and many others. These programs make use of the EM algorithm to optimize the probability of finding the sample of observed genotypes by calculating the frequency of possible haplotypes (Excoffier and Slatkin 1995; Fallin and Schork 2000). Each genotype is thus assigned the haplotype pair with the highest frequency amongst all other haplotype pairs. The EM algorithm is arguably the most popular statistical algorithm of all three methods, because it is easy to interpret, is fairly stable, and is based on well-established statistical properties (Qin et al. 2002). Its performance is also not much affected by deviation from HWE although the EM algorithm makes an explicit assumption of HWE (Niu et al. 2002; Niu 2004). One of the caveats in using the EM algorithm is that the results are dependent on the initial value of haplotype frequencies being reasonably close to the true population frequencies. The existence of local

maxima can result in a false maximum-likelihood estimate and this problem can be elevated as the number of polymorphisms increases (Niu 2004). Most EM-based approaches are best restricted to analysis of less than twenty loci (Zhang et al. 2005), because they are too computationally intensive and cannot handle very large numbers of potential haplotypes. The partition-ligation (PL) algorithm as in the PL-EM software [http://www.people.fas.harvard.edu/~junliu/plem/] is a derivative method incorporating Bayesian Monte Carlo statistics into the basic expectation-maximization (EM) model to increase input capacity (Niu et al. 2002). It first divides the region into artificial blocks and uses the EM algorithm to construct haplotypes within each block. The haplotypes are then ligated via the EM algorithm. Another recently developed program, SNPHAP [http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt] adopts a progressive-extension technique to handle large numbers of linked loci. SNPs are added progressively as haplotype frequencies are estimated and low-frequency ones are discarded (Qin et al. 2002). Both the partition-ligation and progression-extension derivative methods give consistent results but handle only biallelic SNPs. Lastly in software programs such as PHASE [http://www.stat.washington.edu/stephens/software.html] and HAPLOTYPER [http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm], the approach is based on a Bayesian framework calculated via the Markov chain–Monte Carlo (MCMC) technique. Both methods implement a Gibbs sampler for construction of a Markov chain for haplotype frequencies. Unknown haplotypes are regarded as unobserved random quantities and are sampled by consideration of the conditional distribution of the genealogy that underlie the genotype data of randomly sampled individuals, as described by coalescence theory (Stephens et al. 2001). These Bayesian-based methods are relatively fast, handle large numbers of loci (in the

hundreds) and also allow missing genotype data. However, as PHASE imposes a coalescent assumption model on the distribution of unknown haplotypes (as opposed to PL-EM), it would perform better than PL-EM when the assumption is valid and worse than PL-EM when the assumption is not (Zhang et al. 2005).

The main disadvantage of using statistical inference programs is that there is a likelihood that a proportion of inferred haplotypes may be incorrect. The uncertainty in constructing haplotypes will lead to a loss of power in testing for an association with a disease and therefore should be kept to as low as possible. Lin et al., 2002 showed that using a modified Stephens-Smith-Donnelly (SSD) algorithm in a Markov chain–Monte Carlo technique, accuracy of haplotype inference is higher using only common polymorphic sites rather than including all loci (Lin et al. 2002). It is reassuring to note that any of EM or Bayesian-based haplotype inference methods performed almost equally well in both real and simulated studies, resolving haplotypes of at least 1% frequency (Adkins 2004; Xu et al. 2004), although Niu highlighted the sensitivity of these programs to haplotype diversity (Niu 2004). There is therefore a high probability that any of these methods would be able to detect a haplotype that carries a disease risk or phenotypic trait in a significant proportion of the population.

In this study, two programs based on EM, Arlequin, and SNPHAP for haplotype inference were chosen. For data sets less than 20 polymorphic sites, Arlequin was implemented. For data sets more than 20 polymorphic sites and only if these sites are biallelic markers, SNPHAP was more suitable. For data sets of approximately 20 polymorphic sites and requiring analysis of trio data, the Bayesian-based program TAGIT [http://popgen.biol.ucl.ac.uk/people/mw/mike_home.html] was more adept at handling multiple SNP loci in a short time.

1.9 Haplotype tagging

Whilst the costs of genotyping have gone down considerably with the advent of high-throughput techniques, the number of markers to analyze for any candidate gene or a specific region of interest is still too overwhelming to perform for the average laboratory. The presence of LD patterns offers a possibility to select a smaller subset of markers by discarding those markers that are in high LD. Haplotype tagging is a means of selecting a set of non-redundant markers from an initial given set of densely spaced SNPs to retain the most information about the dense map. These markers have variously been called "haplotype-tagging SNPs", "htSNPs", "tagSNP", "tagging SNPs", or "tSNPs". Tagging SNPs (tSNPs) are therefore particularly useful in association studies as they explain the majority of LD patterns within the markers and reduce genotyping effort significantly without reducing resolution (Johnson et al. 2001).

Several statistical methods have been developed for the identification of tSNPs. In general, these SNPs can be determined via haplotype-based or linkage disequilibrium-based methods. At present it is unknown which of these methods are truly superior (Halldorsson et al. 2004b). Haplotype-based methods select tSNPs through optimizing resolution of existing haplotypes and are therefore reliant on correctly inferred haplotypes (Zhang et al. 2002; Sebastiani et al. 2003). Furthermore the generation of large numbers of haplotypes from highly-dense sets of markers within most candidate genes and genomic regions mean that the most common haplotype may not have a population frequency of more than 5% (Crawford and Nickerson 2005). Incorrectly inferred and large number of haplotypes may therefore potentially reduce the efficiency and accuracy in which tSNPs can be inferred. The earliest method by Johnson 2001 simply resolves haplotypes generated while recent and more complex

haplotype-based algorithms construct haplotype blocks as part of their algorithm (Zhang et al. 2002; Ke and Cardon 2003). This assumption of haplotype blocks within the genomic area under study may limit haplotype-based algorithms to be performed on genes with simple haplotype architecture. In LD-based algorithms (Ke and Cardon 2003; Stram et al. 2003; Wang and Xu 2003; Weale et al. 2003; Carlson et al. 2004; Halldorsson et al. 2004a), a quality measure is used to define a set of tagging SNPs that can capture the variance observed. Most of these LD-based methods set a minimum arbitrary threshold of a pair-wise LD measure such as r^2>0.85 (Wang and Xu 2003; Weale et al. 2003; Carlson et al. 2004) so that the selected tSNPs resolve the majority of the haplotypes. While some groups elected to select tSNPs based on pair-wise r^2 values with the tagged SNPs, Weale et al., 2003 used a multilocus linkage disequilibrium measure, haplotype r^2, as a metric of information to select minimum informative subsets of tSNPs (Goldstein et al. 2003; Weale et al. 2003). Using a simplistic LD measure as r^2 is easy to interpret but may be inefficient due to the generation of large numbers of tSNPs (Goldstein et al. 2003). Linkage-disequilibrium-based methods do not rely on haplotype block definition and may therefore be more suitable for genes or genomic regions with complex haplotype architectures. Without prior information about the LD structure inherent in the study, we adopted the use of TAGIT as outlined in Weale's paper for the construction of tSNP sets (Weale et al. 2003).

However several groups have reported that sets of tSNPs differ across different populations (Weale et al. 2003). In studying specific genes that may be involved in predisposition to complex disease or aberrant drug response, LD patterns need to be defined at a finer scale in order to obtain accuracy in candidate gene association

studies (Tishkoff and Verrelli 2003). This suggests that additional studies in specific ethnically diverse populations and individual candidate genes are still necessary.

1.10 International Hapmap Project

The goal of the International Hapmap Project [http://www.hapmap.org/] is to determine the common patterns of DNA sequence variation in the human genome (The International Hapmap Consortium 2003; Deloukas and Bentley 2004). These patterns will be useful in identifying genes that contribute to disease and drug response. Sequence variants will be characterized with their frequencies, and LD data between them. While the most common sequence variations can be obtained in any population, samples from four large populations at different ancestral geographical locations would be included to ensure that allele frequency changes, if any, can be observed amongst populations. Thirty sets of trios from a US Utah population with Northern and Western European ancestry (collected by the Centre d'Etude du Polymorphisme Humain (CEPH)), thirty sets of trios from a Yoruban population in Ibadan, Nigeria, 45 unrelated Japanese in Tokyo, Japan, and 45 unrelated Han Chinese in Beijing, China constitute the 270 samples in this project. To adequately describe the genetic variation across the entire genome, SNP density must be high. The initial intention of the project was to genotype successfully 600,000 SNPs spaced at approximately 5-kilobase intervals, each having a minor allele frequency of at least 5%. Choice of SNPs was to be based on previous validation and distance between consecutive markers. Associations between these SNPs are to be analyzed and in genomic regions where associations are weak, additional SNPs will then be genotyped to increase the SNP density. Today, the number of validated SNPs has risen to more than a million in each of the 4 populations. Still, there are several spare regions where

SNP coverage is still poor. As a result of the Hapmap Project, cross-validation of SNPs from different genotyping platforms results in better corroboration of data. Five genotyping platforms are used: MassExtend (Sequenom, Invader (Third Wave), Acycloprime-FP (Perkin-Elmer), Golden Gate-BeadArray (Illumina) and ParAllele BioScience (Deloukas and Bentley 2004) and may pave the way for the use of large-scale genotyping for future genetic studies (Deloukas and Bentley 2004). The International Hapmap Project holds much promise as a publicly available resource to enable investigators around the world to discover genetic factors that contribute to disease susceptibility or therapeutic response. This scientifically ambitious project will supplement the Human Genome Project and is more than just a catalogue of common genetic variants. The determination of common haplotypes generated from the LD map of the genome serves as a powerful tool for pharmacogenetic and pharmacogenomic studies. Firstly, candidate gene approaches can rely on this resource to look for rare variants that are not genotyped but in LD with underlying common haplotypes found. Secondly, data generated can be extrapolated to common complex diseases throughout the world for populations with similar SNP associations. For populations with ancestry and geographical locations clearly distinct from the samples used in the Project, its data would still serve a basis for comparison. Whole genome–based approaches are then able to seek new targets of functional importance without knowledge of biochemical pathways and gene interaction a priori. Lastly, the Hapmap also drives the generation of tagging SNPs in the identification of human variants that alter function. Other tools based on present knowledge of demographic history of different human populations to screen such potential variants for further studies would also greatly reduce the amount of work and time.

1.11 Detection of positive selection

Various forces such as genetic drift, mutation, recombination and natural selection can influence patterns of genetic diversity. Of particular interest to population and evolutionary geneticists is the inference of past selection events (Altshuler and Clark 2005). Genes under the influence of recent positive selection can be reasonably assumed to evolve as a result of local adaptation to environmental changes (Altshuler and Clark 2005). These changes may be the cause of interindividual variability to disease susceptibility and drug response seen in modern humans. For example, variants that increase the 'fitness' of the population will increase in frequency over time while deleterious mutations at functionally important sites will decrease in frequency (Rebbeck et al. 2004). Researchers have begun looking at patterns of natural selection in the human genome for the purpose of correlating genes to human traits (Fay et al. 2001; Akey et al. 2004). As polymorphisms are the footprints of past genetic events, many methods have been designed to track selection through polymorphisms. Various aspects using polymorphism data can be used to test for selection, including allelic frequency spectrum, nucleotide diversity, polymorphism/divergence, linkage disequilibrium and comparison of the rates of substitution dN/dS (Table 1). Most tests detect selection by rejecting the assumption of neutrality, i.e. the observed data deviates significantly from what is expected under neutrality. Therefore the null model is one in which natural selection is absent. It should be noted that a deviation from neutrality could also be due to other demographic factors such as changes in population size or genetic drift (Przeworski 2002).

Comparison of the rates of substitution: the dN/dS ratio tests

The presence of positive selection can be detected by simply comparing the number of nonsynonymous (dN) to the number of synonymous (dS) substitutions in a locus. Most nonsynonymous substitutions would be expected to be eliminated by purifying selection, but some might be retained under positive selection. Investigating the number of synonymous and non-synonymous substitutions may therefore provide information about the degree of selection operating on a system. Under neutral evolution dN/dS would be equivalent to 1. A high (>1) dN/dS value suggests fixation of nonsynonymous mutations with a higher probability than neutral (synonymous) ones. The dN/dS ratio tests also take into account of other factors such as transition/transversion rate bias and codon usage bias.

Neutrality Test	Brief Description	Strengths / limitations
Frequency-distribution tests		
Ewens-Watterson Homozygosity test	Uses the Ewens sampling formula to compare the observed homozygosity with the expected homozygosity and if the difference between the two homozygosity values is larger than a critical value, the neutral null hypothesis can be rejected.	Vulnerable to deviation from allele frequency distribution equilibrium and is limited in power to detect natural selection.
Tajima D statistic	Compares the difference between the number of segregating sites (S) and the average number of SNPs between sequences in a sample (π). Under the neutral model, *D* value will be negative under selective sweeps (and population growth) and positive under balancing selection (or population substructure).	Tajima's D and Fu & Li's statistics are based on allelic variation and results may not clearly distinguish between selection and demographic alternatives (bottleneck, population subdivision). For example, a large negative Tajima's D value may be obtained if there is a recent selective sweep, a recent population expansion possibly out of bottleneck or the sample is made out of individuals from several smaller subpopulations. On the other hand, a large positive Tajima's D value can be caused by old balancing selection, recent population reduction or recent admixture of two or more highly divergent subpopulations. Both tests are robust to recombination.
Fu and Li D statistic	Compares the number of derived mutations and the number of ancestral mutations. Under a model of neutrality, the frequencies of the derived and ancestral mutations should be the same. Under balancing selection, the variation is maintained longer then expected under a neutral model and results in more ancestral SNPs, thereby giving a positive D value. Under positive selection, there are more derived SNP branches, thereby giving negative D values.	
Fay and Wu H statistic	Compares the difference between the frequency distribution of derived polymorphism in a sample and a unique distribution of polymorphism that is primarily sensitive to the positive selection and not population expansion. Under neutrality, the two test statistics should be close to 0. In the presence of excess high frequency alleles relative to neutral model, H is negative (Przeworski 2002). (http://www.genetics.wustl.edu/jflab/htest.html)	Designed to specifically detect positive selection in the presence of recombination.
Haplotype Tests		
Haplotype number	Detects whether there is a higher or lower number of haplotypes than is expected given the number of segregating sites. When only a few distinct haplotypes are obtained, balancing selection or population structure might have occurred. On the other hand excessive number of haplotypes might be indicative of selective sweeps or population growth.	Unlike frequency-distribution tests, haplotype tests uses information not only from the frequencies of the segregating sites, but also from the haplotype structure. The 3 haplotype tests are correlated. A subset with low number of haplotypes tends to have low haplotype diversity and thus be of a "strong haplotype structure".
Haplotype diversity (H)	Similar to the Watterson's Homozygosity test, but also uses information derived from the number of haplotypes. Under balancing selection, haplotype diversity is high. Under selective sweeps or population growth, haplotype diversity is low.	
Haplotype partition (HP)	Designed to detect a reduction in variation in a subset of sample as compared to the total sample. This could be caused by an incomplete selective sweep.	Relies on observation of a strong haplotype structure for a small subset of the sample.
Tests for heterogeneity		
The Hudson, Kreitman & Aguade (HKA) test	Compares the ratio of silent polymorphism and fixation at a gene region of interest with the ratio of silent polymorphism and fixation at a gene region that is considered neutrally evolving (i.e. the test locus). If both are neutral, the ratio should be similar. Selection in either locus will cause the variance of the ratio to increase.	Most well-known heterogeneity test in population genetics. Less sensitive to demographic changes than frequency-distribution tests.
The McDonald-Kreitman test	Relies on comparing the ratio of synonymous/silent and nonsysnonymous/replacement polymorphisms with the ratio of synonymous and nonsysnonymous fixations within a single gene. Silent polymorphism and fixation reflect the neutral mutation rate (i.e. 4Neμ). An excess of synonymous polymorphisms and fixations indicates adaptative evolution. An excess of nonsynonymous polymorphisms indicate purifying selection while an excess of nonsynonymous fixations represent balancing selection.	However by pooling polymorphisms in categories, this test (as well as the HKA test) does not take into account of the frequency distribution of the segregating polymorphisms and may therefore lose some power in detecting selection (Akashi 1999; Williamson et al. 2005).
The Lewontin-Krakauer test	Compares the observed variance of the Wright's fixation index Fst among loci against the theoretical variance under a model of geographical structure alone (Lewontin and Krakauer 1973). This is based on the two assumptions that gene frequency of each subpopulation is a random sample from a given frequency distribution and that Fst is the same for all loci. If the ratio of observed to expected variance is significantly large, the hypothesis of neutral evolution at all loci in the sample must be rejected. That infers that at least one locus within the set of loci is with unusually high Fst and must be a result of a selective sweep.	The validity of Lewontin-Krakauer test as a test of natural selection was called into question when it was discovered that several processes such as phylogenetic history and migration could inflate the variance in Fst (Robertson 1975b; Robertson 1975a). Despite this, Fst has continued to be a popular choice of measuring variation in allele frequencies among populations (Fst) as an indirect estimator of gene flow and selection (see section under New tests of selection).

Table 1. Summary of the major classes of tests for natural selection. (Adapted primarily from McVean [http://www.stats.ox.ac.uk/~mcvean/L4notes.pdf]) ((Lewontin and Krakauer 1973; Lewontin 1974; Robertson 1975b; Robertson 1975a; Akashi 1999; Kreitman 2000; Przeworski 2002; Williamson et al. 2005)

1.12 New tests of selection

Wright's fixation index

In addition to the previous traditional tests described, new strategies of detecting positive selection have been developed. Akey et al., 2002 examined variation in allele frequency using the F_{st} statistic by comparing the same loci in different test populations instead of amongst different loci (Akey et al. 2002). The Wright's fixation Index F_{st} is the variance of the allele frequencies and is essentially a measure of population subdivision. It makes a comparison of the mean heterozygosity averaged over 2 or more test subpopulations and the heterozygosity calculated if these populations were to be pooled into a single pool. Simply put, F_{st} = (Ht - meanHs) / Ht, where Ht = 2 (pooled p) (pooled q) [note that pooled q = (1-pooled p)], and meanHs = the average of H values for each of the individual subpopulations (i.e., 2 p q for each subpopulation, averaged across all subpopulations).

Under selective neutrality, F_{st} is null and populations have similar allele frequencies implying that there is no differentiation among populations due to genetic drift and therefore the populations are genetically indistinguishable. As F_{st} increases toward 1, genetic drift is causing heterozygotes to be lost within populations, and populations are being fixed for different alleles. Previous tests using F_{st} such as the Lewontin-Krakauer test rely on simulations to obtain expected F_{st} distribution under selective neutrality, but this method relies on prior assumptions on the population demographic history. To avoid this caveat and tap into the tremendous SNP information from The SNP Consortium, Akey et al., 2002 performed an analysis on more than 26000 SNPs in populations of African American, East Asian and European American descent and simply identified genes or loci under natural selection as extreme outliers in the F_{st} distribution (Akey et al. 2002).

Long-Range Haplotype (LRH) test

Sabeti et al., 2002 introduced a novel method that relied on the length of linkage disequilibrium on one haplotype in relation to the frequency of that haplotype (Sabeti et al. 2002). Under neutral evolution, it is expected that the time taken for a newly introduced mutation to rise in allele frequency and accumulate in a population will be long. Due to recombination surrounding the variation, the LD between this variation and nearby variations will decay substantially during this time. If this newly introduced mutation were to be subjected to positive selection, its allele frequency would increase in such a short time that recombination did not substantially break down the haplotype on which the mutation occurred. Therefore an allele with unusually long-range haplotype (and therefore long flanking stretches of homozygosity) given its allele frequency would be an indication of recent positive selection. This method, named the Long-Range Haplotype (LRH) test examines multi-allelic associations using a statistic called relative extended haplotype homozygosity (rEHH), which is in turn derived from extended haplotype homozygosity (EHH). Haplotype homozygosity is the multilocus measure of linkage disequilibrium that calculates the probability of a particular haplotype in a population (Sabatti and Risch 2002). EHH is defined as the probability that two randomly selected chromosomes carrying the core haplotype of interest are identical by descent from the entire interval between the core locus to a distant point x. In simpler terms, it means that EHH represents the probability of two chromosomes that share the same core gene haplotype also share identical haplotypes in other SNPs surrounding the specified candidate core region (Brookfield 2003). As distance increases from the core region, it is expected that EHH will decrease and this rate of decay can be compared amongst core gene haplotypes. rEHH therefore is a comparison of EHH

values of the core haplotype of interest against other core haplotypes and can also be considered to be a measure of extended linkage disequilibrium. A core haplotype with both high frequency and high rEHH, as compared to another with a faster rate of decay in EHH for the same distance, would then be indicative of recent positive selection (Figure 2 Panel A). In Sabeti's paper, he tested core haplotypes within different genes against markers at increasing distance away from the test core haplotype (Sabeti et al. 2002) (Figure 2 Panel B). The advantage of using this test is apparent. Since it compares core haplotypes against one another, it is less vulnerable to demographic changes, specifically the local recombination rate. Sabeti et al., 2002 compared this method against traditional tests of selection (including Tajima's D-test, Fu and Li's D-test, Fay and Wu's H-test, the Ka/Ks test, the McDonald-Kreitman test, and the McDonald-Kreitman-Aguade (HKA)) test in 3 genes, glucose-6-phosphate dehydrogenase (G6PD), and CD40 ligand gene (TNFSF5) and showed this test has greater power in detecting selection. The extended HH, with variances and core haplotype frequencies can be graphically represented using a web-based tool [http://ihg.gsf.de/cgi-bin/mueller/webehh.pl] (Mueller and Andreoli 2004). Nonetheless, this test has its own limitations. Firstly it detects only recent positive selection. For changes in allele frequency driven by selection early on in human history, recombination would have time to dissolve any linkage disequilibrium in flanking sequences (Brookfield 2003). Secondly it is also dependent on the assumptions that the human population is constant in size and contains no substructure (Brookfield 2003). Thirdly it might be subjected to sampling variations, deriving different haplotypes and haplotype frequencies.

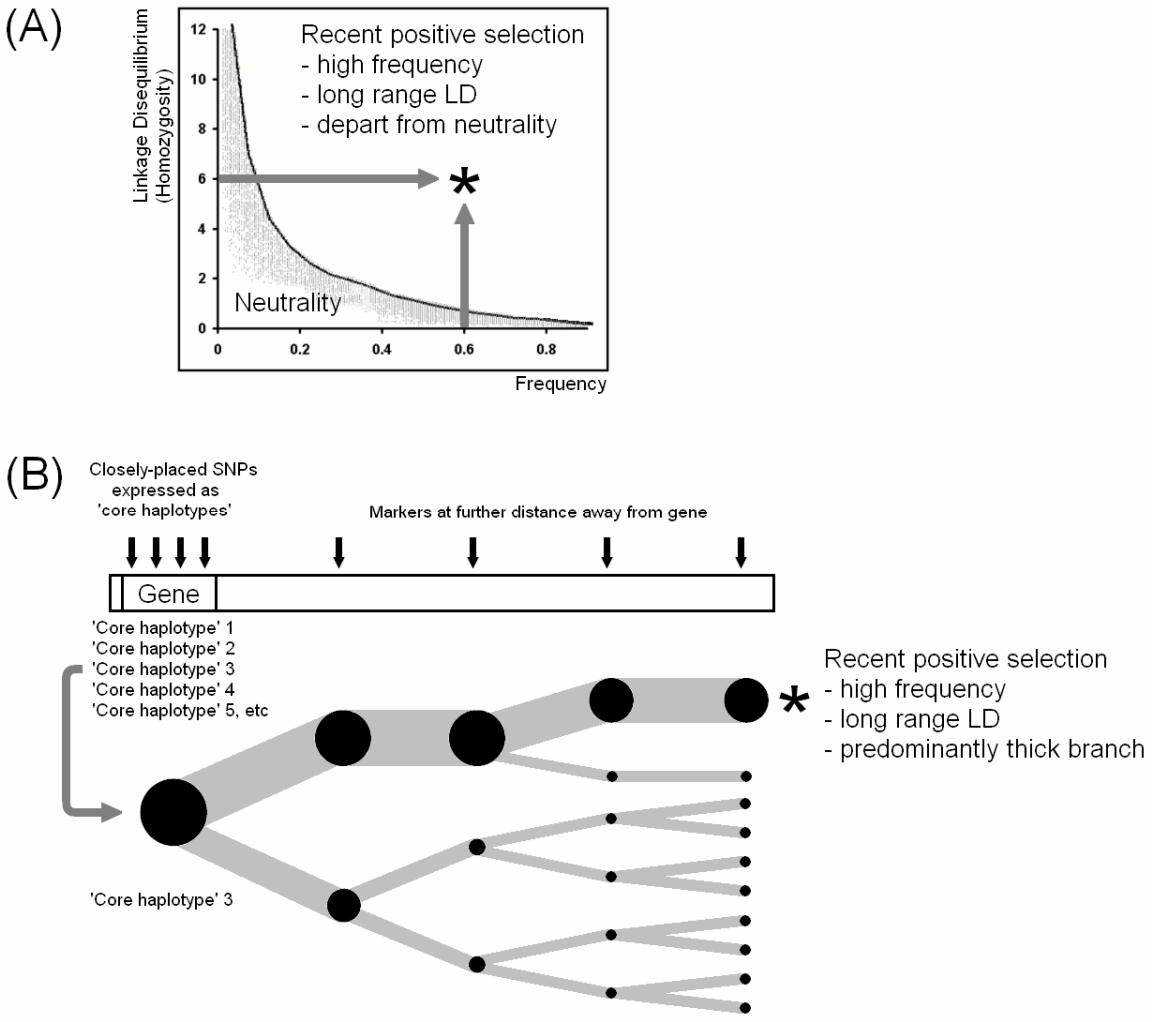


Figure 2. Long-range haplotype test.
Panel A: Extended haplotype homozygosity (EHH) as an LD measure. A high-frequency haplotype that possesses long-range LD departs significantly from neutrality and thereby shows evidence of recent positive selection. Various coalescent models e.g. of constant size, expansion, extreme bottleneck and highly structured populations are simulated to test haplotypes for other demographic influences. Panel B: Haplotype Branching Diagram (HBD). Closely-placed SNPs within genes are expressed as 'core haplotypes'. Each of the core haplotypes serves as a core locus from which branches signifying the profile of LD decay appear. As LD decays from this core locus to consecutive markers placed further away, more branches appear. The thickness of each branch corresponds to the proportion of chromosomes within the haplotype. A predominantly thick LD branch is evidence of recent positive selection.

Pexcess test statistic

Bersaglieri et al., analyzed the *LCT* gene, which encodes the enzyme lactase-phlorizin hydrolase in multiple populations for signature of positive selection using 3 different measures (Bersaglieri et al. 2004). In addition to the F_{st} and LRH tests, they described a novel test using the *P*excess statistic. This test calculates the probability P which represents the extent at which a haplotype is over-represented in the affected sample, and is analogous to **P**excess = (**P***test* - **P***reference*)/(1 - **P***reference*), where **P***test* and **P***reference* denote allele frequency in chromosomes of population under study (test) and ancestral (reference) population, respectively. The ancestral allele frequency (**P***reference*) is estimated by taking the average allele frequency in populations that have not experienced selection such as that of East Asian and African American populations (Bersaglieri et al. 2004). By observing consistently elevated *P*excess values that extend across multiple markers in a large region (>50kb), the increase in frequency of a single haplotype can be approximated. Looking for positive selection using elevated *P*excess values across multiple loci may be more informative than elevated F_{st} values in individual markers. However the Pexcess test is dependent on the assumptions of all allele frequency differentiation measures. In addition, using allele frequency of the African American or East Asian population as the ancestral allele frequency may be erroneous, even though Bersaglieri et al., showed that their results did not defer significantly whether 21% European admixture in the African American population was accounted for (Bersaglieri et al. 2004).

Chapter 2 The nucleotide analogue transporters

2.1 Drug transporters

While whole genome scanning is likely to be widely used in the future, candidate gene studies are, at present, more feasible. One of the key issues in performing a candidate gene case-control association study is the selection of candidate genes. Justifications are required for the choice of genes or genomic regions in which polymorphisms are to be analyzed. The number of genes or genomic regions may be numerous for complex diseases and drug response, hence it is important to prioritize these genes so that variations in them can be sequentially examined. Furthermore with the ever-expanding number of polymorphic sites discovered in disease- or drug-related genes, it is important that the functional significance of the polymorphisms chosen for study and their reported allele frequencies are taken into account so that genuine associations are detected.

Genetic variability can affect drug response in two ways, pharmacodynamically and pharmacokinetically. Most drugs act by interacting with target proteins such as receptors and if polymorphisms change the protein configuration of the target protein, drug response may be affected. Any genetic variability in genes involved in any one of the ADME (Absorption, Distribution, Metabolism, and Excretion) functions may also affect drug levels at target sites. Drug transporters are membrane bound proteins that actively export therapeutic drugs out of cells and are thus able to influence the pharmacokinetic characteristics of drugs (absorption, distribution, excretion). These transporters are also able to influence drug penetration into the brain and central nervous system. Some of them are also implicated as a mechanism of cancer multidrug resistance and linked to some genetic diseases. The energy-dependent efflux of drug from cells is a critical concept that first arose when a glycosylated

plasma membrane protein was detected in cells resistant to colchicine (Juliano and Ling 1976). The protein P-glycoprotein (P-gp), encoded by the MDR1 gene, was subsequently shown to impart resistance to a wide array of structurally and chemically unrelated compounds. Since then, other members of the ATP-binding cassette (ABC) superfamily of transporters have been identified. The multidrug resistance related proteins (MRPs), belonging to the *ABCC* subfamily, hold special interest as substrates range from a simple ion (*ABCC*7/MRP7/"cystic fibrosis transmembrane conductance regulator"/CFTR) to various anionic compounds (including large drug molecules) and have been linked to pathological states such as pseudoxanthoma elasticum (*ABCC*6/MRP6) and cystic fibrosis (*ABCC*7/MRP7). The analysis of the predicted structures show that *ABCC*4/MRP4, *ABCC*5/MRP5, *ABCC*11/MRP8 and *ABCC*12/MRP9 share structural similarity to MDR1 in having two nucleotide binding domains (NBDs) and two membrane spanning domains (MSDs), each composed of six transmembrane helices with a linker region (L1) connecting NBD1 to MSD2 (Figure 3 Panel A) (Gottesman et al. 2002; Kruh and Belinsky 2003). The basic structure of other subfamily members such as *ABCC*1/MRP1, is composed of the same MDR1 core region, but has an additional N-terminal region composing of extracellular N-terminus, a membrane spanning domain MSD_0, and an intracellular loop (L_0) connecting MSD_0 to MSD_1 (Kruh and Belinsky 2003) (Figure 3 Panel B). The two cytosolic NBDs of about two hundred amino acids contain ATP-binding sites with three consensus sequences conserved among species (Efferth 2001; Gottesman and Ambudkar 2001; Lockhart et al. 2003). The Walker motifs A and B are involved in ATP binding and hydrolysis (Nikaido 2002), and are linked by a third consensus motif. This unique motif is called LSGGQ motif, C motif, or “signature motif S” as it is present in all ABC transporters (Figure 4) (Klein et al. 1999; Bodo et al. 2003). In

terms of amino acid identity, *ABCC*1/MRP1, *ABCC*2/MRP2, *ABCC*3/MRP3 and *ABCC*6/MRP6 proteins share 45-58% similarity whereas both the *ABCC*4/MRP4 and *ABCC*5/MRP5 share less similarity with *ABCC*1/MRP1 (36-39%) (Lee 2000) (Table 2). As yet, no work has been published on the characterization of the promoter regulatory elements of the *ABCC*4 and *ABCC*5 loci (Bush and Li 2002).

		Overall percent amino acid identity					
Protein	**Amino**	**ABCC1**	**ABCC2**	**ABCC3**	**ABCC4**	**ABCC5**	**ABCC6**
ABCC1	1531 aa	-					
ABCC2	1545 aa	48.4	-				
ABCC3	1527 aa	57.6	46.8	-			
ABCC4	1325 aa	39.4	36.8	35.3	-		
ABCC5	1438 aa	35.8	36.2	33.1	36.5	-	
ABCC6	1503 aa	45.0	39.1	43.6	33.9	30.9	-

Table 2. Overall amino acid identity amongst selected *ABCC* subfamily members. (Adapted from Belinsky et al. 1998; Kool et al. 1999).

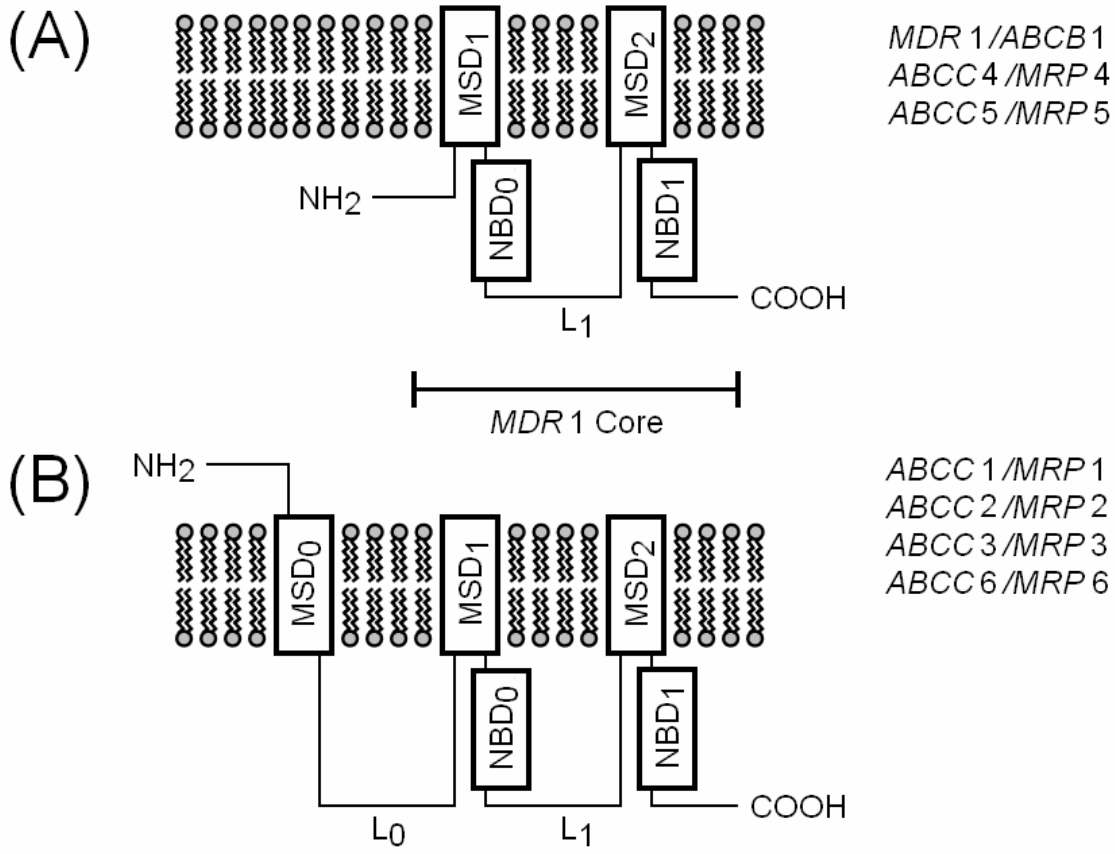


Figure 3. Structural illustration of selected members of ABC transporter superfamily. [Adapted from (Gottesman et al. 2002)] Panel A: Structural similarity of *ABCC*4/MRP4 and *ABCC*5/MRP5 to MDR1. Panel B: *ABCC*1, *ABCC*2, *ABCC*3 and *ABCC*6 possess an additional N-terminal region in addition to the MDR1 core region.

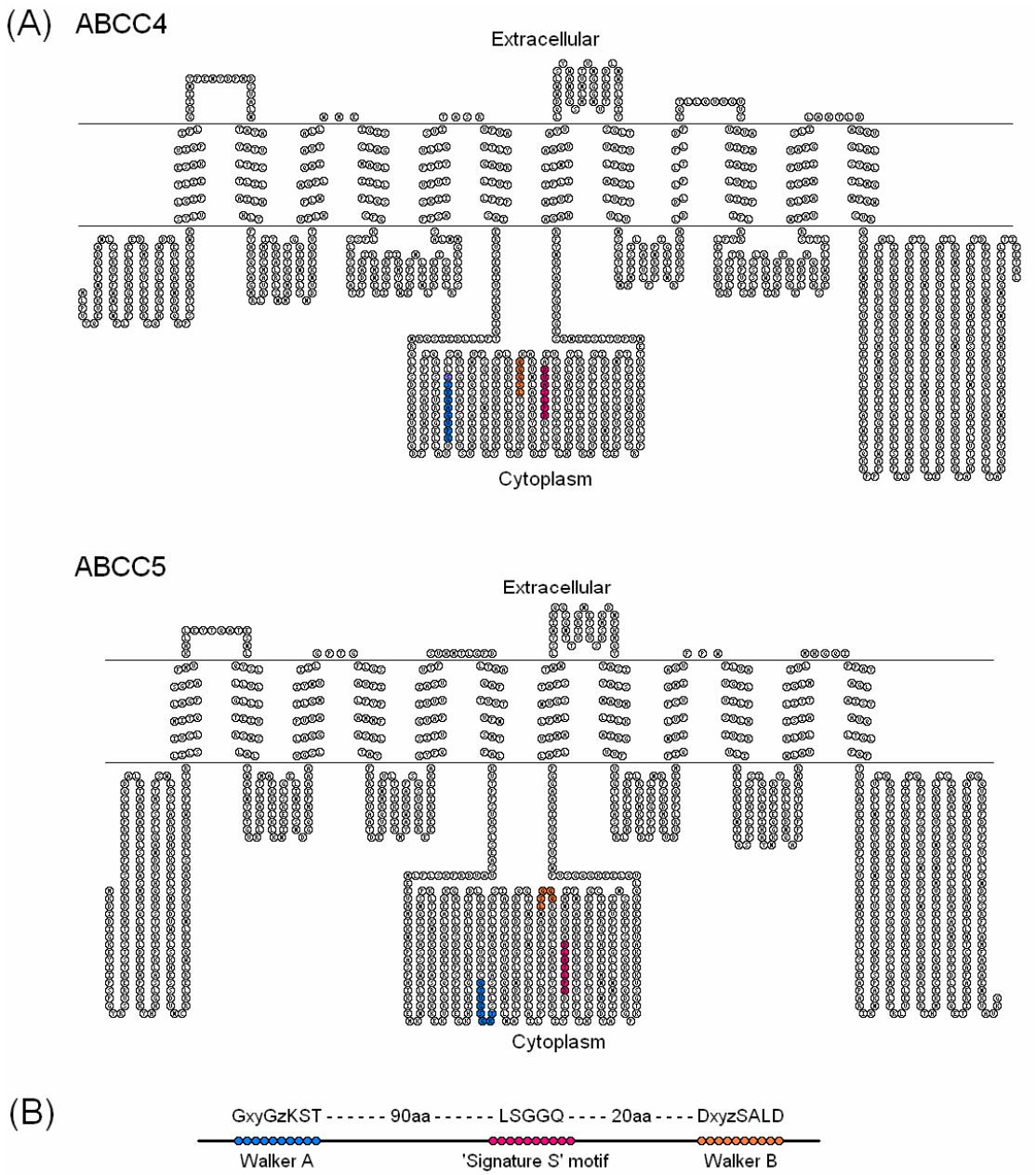


Figure 4. Predicted secondary structures of *ABCC*4/MRP4 and *ABCC*5/MRP5 proteins. The transmembrane topology schematic was rendered using a transmembrane protein display software available at TOPO2 [http://www.sacs.ucsf.edu/TOPO-run/wtopo.pl]. Panel A: Prediction of transmembrane regions in *ABCC*4/MRP4 and *ABCC*5/MRP5 using TMHMM Server v. 2.0 [http://www.cbs.dtu.dk/services/TMHMM-2.0/] and "DAS" - Transmembrane Prediction server [http://www.sbc.su.se/~miklos/DAS/tmdas.cgi]. Panel B: Characteristic motifs within *ABCC*4/MRP4 and *ABCC*5/MRP5 proteins. Walker A and Walker B regions are linked by a 'signature motif S' (Gottesman and Ambudkar 2001). These regions are also highlighted in the respective protein structures in Panel A.

2.2 *ABCC*4 and *ABCC*5 in antiretroviral and anticancer therapy

*ABCC*4, *ABCC5* and *ABCC*8 proteins are the only MRP isoforms shown to confer resistance to cyclic nucleotides (cAMP and cGMP), acyclic nucleoside phosphonates such as 9-(2-phosphonylmethoxyethyl)adenine (PMEA), and monophosphorylated nucleoside analogues such as azidothymidine-monophosphate (AZT-MP; 2'-azido-2',3'-dideoxythymidine monophosphate, zidovudine-monophosphate), thioxanthosine monophosphate, and thioinosine monophosphate (Schuetz et al. 1999; Jedlitschky et al. 2000; Wijnholds et al. 2000; Chen et al. 2002; Lai and Tan 2002; Wielinga et al. 2002; Guo et al. 2003). The multidrug resistance protein 5 (MRP5/*ABCC*5) was the first ATP-dependent transporter identified for cyclic nucleotides with cGMP as a high-affinity substrate and cAMP as a low-affinity substrate (Jedlitschky et al. 2000) (Table 3). While *ABCC*4 has been recently shown to mediate efflux of cyclic nucleotides (van Aubel et al. 2002), its affinity for cGMP is lower than that of *ABCC*5 (Adachi et al. 2002; Chen et al. 2002; Lai and Tan 2002) (Table 3). Despite these findings, Klokouzas et al's (2003) vesicle uptake studies using MRP inhibitors suggested that the properties of a cGMP transporter in human erythrocytes was similar to those of the *ABCC*4 transporter, and not those of the ABCC5 transporter (Klokouzas et al. 2003). Interestingly, the affinity of cAMP for *ABCC*4 is at least 9 fold higher than that for *ABCC5*.

Nucleoside analogues are useful for both antitumour and antiviral therapies. Once taken up by cells, these analogues must be activated by phosphorylation, and incorporated into nucleic acids before exerting their cellular toxicity (Reid et al. 2003a). Both *ABCC*4 and *ABCC*5 transporters were shown to confer similar levels of resistance to unmodified PMEA, but not its phosphorylated form (Lee et al. 2000; Wijnholds et al. 2000; Lai and Tan 2002; Reid et al. 2003a). It was noted that

transport of both efflux pumps was limited to nucleoside monophosphates, but not the base, nucleoside, or di-phosphonate or tri-phosphonate derivatives of the nucleoside (Reid et al. 2003a; Wielinga et al. 2003; Borst et al. 2004). *ABCC*4 and *ABCC*5 proteins do not confer resistance against typical substrates of *ABCC*1-3 anthracyclines (such as daunorubicin and doxorubicin), etoposide, vincristine, paclitaxel, cisplatin, heavy metals (cadmium chloride, potassium antimony tartrate and sodium meta-arsenite) (McAleer et al. 1999; Lee et al. 2000; Wijnholds et al. 2000) and are therefore functionally distinct from these proteins. In fact, negative correlation of *ABCC*5 mRNA levels with the resistance of unselected lung cancer cell lines to etoposide (VP-16) and vincristine (VCR) had been reported (Young et al. 1999).

High-level resistance against other acyclic nucleoside phosphonate drugs, such as cPr-PMEDAP [cyclopropyl-PMEDAP] and PMEDAP [9-(2-phosphonomethoxyethyl)-2,6-diaminopurine] was also observed for *ABCC*4 but not *ABCC*5 (Reid et al. 2003a) (Table 3). Low levels of resistance against purine-based carbocyclic nucleoside analogue abacavir (Reid et al. 2003a) as well as significant levels of thioguanine and 6-Mercaptopurine have been observed for both transporters (Wijnholds et al. 2000; Reid et al. 2003a; Wielinga et al. 2003). Some controversies regarding other substrates exist in literature. Chen et al., 2001 reported *ABCC*4 overexpression in NIH3T3 cells does not affect levels of cladribine (Chen et al. 2001), while Reid et al., showed some low levels of resistance by HEK293 cells expressing *ABCC*4 or *ABCC*5 (Reid et al. 2003a). Davidson et al., 2002 reported substantial resistance to cytarabine, gemcitabine, and cladribine in HEK293 cells with overexpression of *ABCC*5 (Davidson et al. 2002b; Ritter et al. 2005) while Reid et al., 2003 could not observed a significant level of resistance against gemcitabine, cytarabine, and fludarabine in HEK293 cells overexpressing either *ABCC*4 or *ABCC*5 (Reid et al. 2003a). Whereas

Jedlitschky et al., 2000 reported that the *ABCC5*-mediated vesicular transport of cGMP was sensitive to inhibitors of cGMP phosphodiesterase 5 such as trequinsin (Jedlitschky et al. 2000), Reid et al., 2003 observed only low levels of inhibition by sildenafil, trequisin, and zaprinast on MRP4- and MRP5-mediated PMEA efflux (Reid et al. 2003a). The repertoire of drugs transported by *ABCC*4 was also recently expanded when it was shown that overexpression of *ABCC*4 restricted accumulation of the topoisomerase inhibitor topotecan via vesicular transport assays and concentration of topotecan in the cerebrospinal fluid (CSF) of MRP4-/- mice was almost 10-fold that of MRP4+/+ mice (Leggas et al. 2004). In another publication by Norris et al., 2005, the *in vitro* accumulation of other topoisomerase inhibitors, namely irinotecan/CPT-11, and its metabolite, SN-38 in *ABCC*4-transduced cells but not *ABCC5*-transduced cells also proved that the expression of *ABCC*4 protein might have implications for the clinical use of camptothecin-based analogues which are now in clinical trials for the treatment of neuroblastoma (Norris et al. 2005). Norris et al., 2005, were, however, unable to observed any significant resistance to either camptothecin or topotecan (Norris et al. 2005).

While the substrate profiles of *ABCC*4 and *ABCC*5 for nucleoside-based molecules are remarkably similar, *ABCC*4 also has the ability to transport conjugates such as the estradiol glucuronide $E_2$17ßG and antimetabolite methotrexate (Lee et al. 2000; Chen et al. 2001; Chen et al. 2002; van Aubel et al. 2002). Other physiological substrates have also been found for *ABCC*4 such as steroid-glucuronides and sulfated steroid (Zelcer et al. 2003), monoanionic bile acid conjugates (in cotransport with reduced glutathione GSH) (Rius et al. 2003), dianionic bile acid conjugates (in absence of GSH) (Zelcer et al. 2003), prostaglandins PGE_1 and PGE_2 (Reid et al. 2003b) as well as urate at low affinities (van Aubel et al. 2005) (Table 3). The *ABCC*4 specific

bindings of p-aminohippurate and purine end metabolite urate as well as its transcript and protein expression on the brush border membrane of renal tubules suggest an excretory role for *ABCC*4 (Smeets et al. 2004; van Aubel et al. 2005). No physiological substrate has yet been found for *ABCC*5. Furthermore, cGMP transport by recombinant MRP4 can be inhibited by substances such as dipyridamole and the leukotriene analogue MK571 as well as the non-steroidal anti-inflammatory drugs (NSAIDs) such as indomethacin (Reid et al. 2003a; Reid et al. 2003b) while *ABCC*5-mediated transport is inhibited by dipyridamole but not MK571 (Jedlitschky et al. 2000; Reid et al. 2003a) (Table 4). Many of these inhibitors are assumed to be substrates as well (Table 4).

ABCC4			*ABCC5*		
Substrate	**K_m**	**Reference**	**Substrate**	**K_m**	**Reference**
Cyclic nucleotides			*Cyclic nucleotides*		
cAMP	44.5 uM 100 uM	Chen et al., 2001 van Aubel et al., 2002 Weilinga et al., 2003 Lai and Tan, 2002 Sampath et al., 2002	cAMP	379 uM	Jedlitschky et al., 2000
cGMP	9.69 uM 1 uM	Chen et al., 2001 van Aubel et al., 2002 Weilinga et al., 2003	cGMP	2.1 uM	Jedlitschky et al., 2000
			8-Bromo-cGMP		Jedlitschky et al., 2000
			N^2,2'-*O*-Dibutyryl-cGMP		Jedlitschky et al., 2000
Nucleoside analogues			*Nucleoside analogues*		
PMEA		Reid et al., 2003a Schuetz et al., 1999 Sampath et al., 2002 Lee et al., 2000	PMEA		Wijnholds et al., 2000
PMEG		Reid et al., 2003a Schuetz et al., 1999	PMEG		Reid et al., 2003a
PMEDAP		Sampath et al., 2002 Reid et al., 2003a	PMEDAP		Reid et al., 2003a
cPr-PMEDAP		Reid et al., 2003a	cPr-PMEDAP		Reid et al., 2003a
bis-POM-PMEA		Reid et al., 2003a	bis-POM-PMEA		Reid et al., 2003a
(S)-HPMPC		Reid et al., 2003a	(S)-HPMPC		Reid et al., 2003a
Ganciclovir		Adachi et al., 2002			
AZT		Schuetz et al., 1999			
3TC		Schuetz et al., 1999			
			Abacavir		Reid et al., 2003a
			Cladribine		Reid et al., 2003a
			Cladarabine		Davidson et al., 2002*
			Gemcitabine		Davidson et al., 2002*
			Cytarabine		Davidson et al., 2002*
			5-HP		Wijnholds et al., 2000
			5-FU	1.3 mM	Dantzig et al., 2003*
5-Br-d-Urd		Sampath et al., 2002			
6-Mercaptopurine (6-MP)		Reid et al., 2003a Chen et al., 2001	6-Mercaptopurine (6-MP)		Wijnholds et al., 2000 Reid et al., 2003a
			Azathioprine (6-MP prodrug)		Wijnholds et al., 2000
6-Thioguanine		Chen et al., 2001 Reid et al., 2003a	Thioguanine		Wijnholds et al., 2000 Dantzig et al., 2003* Reid et al., 2003a
Methotrexate (MTX)	0.22 mM 1.3 mM	Chen et al., 2002 Lee et al., 2000 van Aubel et al., 2002 Sampath et al., 2002			
Ribavirin		Sampath et al., 2002			
Adefovir		Lee et al., 2000 Schuetz et al., 1999			

Table 3. Substrates of *ABCC4* and *ABCC5* proteins.
*Data directly taken from Ritter et al., 2005 (Ritter et al. 2005) with references from Davidson et al., 2002a and Dantzig et al., 2003 (Davidson et al. 2002a; Dantzig et al. 2003; Pratt et al. 2005). ABBREVIATIONS: cAMP, adenosine 3',5'-cyclic monophosphate; cGMP, guanosine 3',5'-cyclic monophosphate; PMEA, 9-(2-

phosphonomethoxyethyl)adenine; PMEG, 9-(2-phosphonomethoxyethyl)guanine; PMEDAP, 9-(2-phosphonomethoxyethyl)-2,6-diaminopurine; cPr-PMEDAP, cyclopropyl-PMEDAP; bis-POM-PMEA, bis(pivaloyloxymethyl)-9-(2-phosphonomethoxyethyl)adenine; (*S*)HPMPC, 1-(*S*)-3-hydroxy-2-(phosphonomethoxy)propyl]cytosine; AZT, zidovudine (azidothymidine); 3TC, 2'-deoxy-3'-thiacytidine (lamivudine); 5-Br-dUrd, 5-bromo-2'deoxyuridine; adefovir [9-(2-phosphonylmethoxyethyl)-[2,8-3H]adenine; $E_2$17βG, estradiol 17-β-D-glucuronide; PG, prostaglandin; DNP-SG, S-(2,4-dinitrophenyl)-glutathione; NAc-DNP-Cys, N-acetyl (2,4,-dinitrophenyl)-cysteine; Leucovorin (folinic acid, N5-formyl-THF, citrovorum factor); 5-HP, 5-hydroxypyridine-2-carboxaldehyde thiosemicarbazone; 5-FU, 5-flurouracil.

***ABCC*4**			***ABCC*5**		
Substrate	**K_m**	**Reference**	**Substrate**	**K_m**	**Reference**
Miscellaneous molecules			*Miscellaneous molecules*		
Glutathione (GSH)	2.7 mM	Rius et al., 2003			
S-methyl-glutathione (Me-SG)	1.2 mM	Rius et al., 2004			
cholyltaurine (C-tau)	3.8 uM	Rius et al., 2005			
$E_2$17betaG	30 uM	Chen et al., 2001			
		Reid et al., 2003b			
		van Aubel et al., 2002			
	30 uM	Zelcer et al., 2003			
DHEAS	2 uM	Zelcer et al., 2003			
Sulfated bile acids		Zelcer et al., 2003			
Prostaglandin PGE_1	2.1 uM	Reid et al., 2003b			
Prostaglandin PGE_2	3.4 uM	Reid et al., 2003b			
Phosphonoformic acid (forscarnet; PFA)		Sampath et al., 2002			
Phosphonoacetic acid (PAA)		Sampath et al., 2002			
S-(2,4-dinitrophenyl)glutathione (DNP-SG)		van Aubel et al., 2002	S-(2,4-dinitrophenyl)glutathione (DNP-SG)		Wijnholds et al., 2000
NAc-DNP-Cys		van Aubel et al., 2002			
Leukotriene C4		van Aubel et al., 2002			
alpha-Naphthyl-beta-D-glucuronide		van Aubel et al., 2002			
para-Nitrophenyl-beta-D-glucuronide		van Aubel et al., 2002			
Folic acid	0.17 mM	Chen et al., 2002			
Leucovorin	0.64 mM	Chen et al., 2002			
			Cadmium chloride ($CdCl_2$)		McAleer et al., 1999
			Potassium antimonyl tartrate		McAleer et al., 1999
					Wijnholds et al., 2000
			5-chloromethylfluorescein diacetate		McAleer et al., 1999
			Fluorescein diacetate		Jedlitschky et al., 2000

Table 3 (continued).

ABCC 4			*ABCC* 5		
Inhibitor	**IC_{50} (uM)**	**Reference**	**Inhibitor**	**IC_{50} (uM)**	**Reference**
Uricosuric agents			*Uricosuric agents*		
Probenecid	<100	van Aubel et al., 2002	Probenecid	~50	Jedlitschky et al., 2000
	~100	Rius et al., 2003		200	Reid et al., 2003a
	2300	Reid et al., 2003a			
Sulfinpyrazone	420	Reid et al., 2003a	Sulfinpyrazone	300	Reid et al., 2003a Wijnholds et al., 2000
Benzbromarone	150	Reid et al., 2003a	Benzbromarone	150	Reid et al., 2003a
Phosphodiesterase 5 (PDE5) Inhibitors			*Phosphodiesterase 5 (PDE5) Inhibitors*		
Zaprinast	250	Reid et al., 2003a	Zaprinast		Jedlitschky et al., 2000
				250	Reid et al., 2003a
Trequensin	10	Reid et al., 2003a	Trequensin		Jedlitschky et al., 2000
				30	Reid et al., 2003a
Sildenafil	20	Reid et al., 2003a	Sildenafil		Jedlitschky et al., 2000
				80	Reid et al., 2003a
Leukotrienes Receptor Antagonists (LTRAs)					
MK571	2	Rius et al., 2003	MK571		Jedlitschky et al., 2000
	10	Reid et al., 2003a		40	Reid et al., 2003a
	43	Jedlitschky et al., 2004			
Nonsteroidal Anti-inflammatory Drugs (NSAIDs)			*Nonsteroidal Anti-inflammatory Drugs (NSAIDs)*		
Dipyridamole	2	Reid et al., 2003a	Dipyridamole	30	Reid et al., 2003a
	12	Jedlitschky et al., 2004			
	<20	Rius et al., 2003			
Indomethacin		Reid et al., 2003a			
	~5	Reid et al., 2003b			
	22	Jedlitschky et al., 2004			
Ibuprofen	20	Reid et al., 2003b			
	~50	Jedlitschky et al., 2004			
Flurbiprofen	~5	Reid et al., 2003b			
Indoprofen	~5	Reid et al., 2003b			
Ketoprofen	~50	Reid et al., 2003b			
Diclofenac	poor	Reid et al., 2003b			
Celecoxib	poor	Reid et al., 2003b			
Rofecoxib	poor	Reid et al., 2003b			
Antiplatelet Drugs			*Antiplatelet Drugs*		
Dilazep	20	Reid et al., 2003a	Dilazep	>50	Reid et al., 2003a
Nucleoside analogues			*Nucleoside analogues*		
Methotrexate		Rius et al., 2003			
Nitrobenzylmercaptopurine riboside	75	Reid et al., 2003a	Nitrobenzylmercaptopurine riboside	>100	Reid et al., 2003a

Table 4. Substrates of *ABCC*4 and *ABCC*5 proteins with inhibitory actions. ABBREVIATIONS: MK571, 3-[[3-[2-(7-chloroquinolin-2-yl)vinyl]phenyl]-(2-dimethylcarbamoylethylsulfanyl)methylsulfanyl] propionic acid

*ABCC*4			*ABCC*5		
Inhibitor	**IC_{50} (uM)**	**Reference**	**Inhibitor**	**IC_{50} (uM)**	**Reference**
Miscellaneous molecules					
Prostaglandin PGF_1alpha		Reid et al., 2003b			
Prostaglandin PGF_2alpha		Reid et al., 2003b			
Prostaglandin PGA_1		Reid et al., 2003b			
Thromboxane TXB_2		Reid et al., 2003b			
Folate		Rius et al., 2003			
Cholate	250	Zelcer et al., 2003			
Glycocholate	400	Zelcer et al., 2003			
Taurocholate	350	Zelcer et al., 2003			
Taurodeoxycholate	60	Zelcer et al., 2003			
Taurochenodeoxycholate	55	Zelcer et al., 2003			
Taurolithocholate	20	Zelcer et al., 2003			
Taurolithocholic acid sulphate	10	Zelcer et al., 2003			
Glycolithocholic acid sulphate	10	Zelcer et al., 2003			
Lithocholic acid sulphate	10	Zelcer et al., 2003			
DHEAS	3	Zelcer et al., 2003			
DHEA 3-glucuronide	80	Zelcer et al., 2003			
Oestrone 3-sulphate	45	Zelcer et al., 2003			
Oestradiol 3-sulphate	50	Zelcer et al., 2003			
Oestradiol 3-glucuronide	120	Zelcer et al., 2003			
Oestradiol 3,17-disulphate	2	Zelcer et al., 2003			

Table 4 (continued).

2.3 Tissue distribution of *ABCC*4 and *ABCC*5 proteins

Expression of *ABCC*4 and *ABCC5* in specific tissues may have an impact on therapeutic outcome by affecting nucleotide accumulation and determine their physiological roles in these tissues. Both nucleotide analogue transporters *ABCC*4 and *ABCC*5 are shown to have a widespread expression in a variety of tissues. *ABCC*4 mRNA has been detected in several tissues: prostate, liver, testis, ovary, brain, kidney, and adrenal gland (Kool et al. 1997; Schuetz et al. 1999; Lee et al. 2000; Zelcer et al. 2003). The expression of *ABCC*4 appears to be the highest in the prostate (Lee et al. 1998), with primary localization in the basolateral membrane of the tubuloacinar cells (Lee et al. 2000). Immunoblotting studies have provided evidence for the expression

of *ABCC*4 and *ABCC*5 in erythrocytes (Klokouzas et al. 2003; Wu et al. 2005), and presence of *ABCC*4 in human platelets (Jedlitschky et al. 2004). As shown by immunofluoresence techniques, *ABCC*4 is localized in the intracellular structures in platelets and plasma membranes in human monocytes (Ritter et al. 2005). *ABCC*4's expression in these tissues might therefore play a role in therapeutic therapy by coordinating the intracellular accumulation of purine nucleosides. For example, the expression of *ABCC*4 has been shown at the apical (brush border) membrane of the human renal proximal tubules and may therefore explain the nephrotoxicity incurred by nucleotide analogue drugs by PMEA, adefovir and cidofovir (van Aubel et al. 2002; van Aubel et al. 2005). The differential polarization of *ABCC4*'s expression in various tissues is thus unique and the biological relevance of such ability to selectively localize either apically or basolaterally is unknown.

The expression of *ABCC*5 mRNA/protein appears to be ubiquitous, with expression shown in many tissues. In a direct comparison between expression of *ABCC*4 and *ABCC*5, Kool et al., 1997 found that *ABCC*5 mRNA was detected in almost all common tissues tested and at much higher levels than *ABCC*4 using RNase protection assays (Kool et al. 1997). Belinsky et al., 1998, used RNA blot analysis to show expression of *ABCC*5 in skeletal muscle (highest), kidney, testis, heart and brain tissues (Belinsky et al. 1998). Spleen, thymus, pancreas, prostate, ovary, placenta, small intestine, colon peripheral blood lymphocytes (PBL), lung and liver also contained low levels of *ABCC*5 (Belinsky et al. 1998). Localization of *ABCC*5 protein in smooth muscle cells as well as endothelial cells of blood vessels in the genitourinary system and cardiomyocytes also suggests that *ABCC*5 plays a regulatory in role in the regulation of vascular smooth muscle tone and cardiac contractility via intracellular cyclic nucleotide concentration as well as in an alternate

mechanism of action for phosphodiesterase inhibitors such as sildenafil (Nies et al. 2002; Dazert et al. 2003).

Of worthy mention with respect to xenobiotic transport is the expression of *ABCC*4 and *ABCC*5 on structural barriers within the human body which restrict entry of foreign materials into certain tissues and organs (Borst and Elferink 2002; Schinkel and Jonker 2003). The blood-brain barrier (BBB) formed by the brain capillary endothelial cells (BCEC) and the blood-cerebrospinal fluid barrier (BCSFB) formed by the choroid plexus form a very effective barrier to the transfer of xenobiotics into the brain (Graff and Pollack 2004).

The presence of both *ABCC*4 and *ABCC*5 is demonstrated in brain microvessel endothelial cells that form the blood–brain barrier (BBB) (Zhang et al. 2000). Kool et al. 1997 also demonstrated a high level of *ABCC*5 expression in brain cell lines (Kool et al. 1997) and other studies have also found high levels of mRNA transcripts within cerebral cortex, cerebellum, and hippocampus (Belinsky et al. 1998; McAleer et al. 1999). Besides endothelial cells, glial cells which consist of oligodendrocytes, astrocytes and microglias are also an important component of the blood-brain barrier and the Central Nervous System (CNS). As such, *ABCC*4 and *ABCC*5 transcriptional expression in glial cells may be important in elucidating the roles of these transporters in BBB physiology. Both *ABCC*4 and *ABCC5* transcripts could be found in BCECs, glial cells, pericytes, capillary extract and grey matter extract (Berezowski et al. 2004). In sharp contrast to the absence of P-gp, the MRP homologue transporters *ABCC*1, 4, 5, and 6 were detected in glial cells (Hirrlinger et al. 2002; Berezowski et al. 2004). As was also shown by Kool et al., 1997, *ABCC*4 mRNA transcription in these cells was much lower than *ABCC*5 (Kool et al. 1997; Berezowski et al. 2004). In addition, Dallas et al., 2004 showed that PMEA efflux from a microglia cell line was mediated

by *ABCC*4 and *ABCC*5 (Dallas et al. 2004). However, it was also shown that there is no involvement of glial cells or pericytes in the regulation of transcriptional expression of ABCC4 and ABCC5 transporters (Berezowski et al. 2004). With studies demonstrating *ABCC*4- and *ABCC*5- mediated resistance to antiviral nucleosides, one of the functions of *ABCC*4 and *ABCC*5 may be involved in nucleoside transport in the BBB. The low CNS penetration of dideoxynucleoside reverse transcriptase inhibitors, such as zidovudine (AZT), may be due in part to presence of efflux transporters including *ABCC*4 and *ABCC*5. The variability of effective therapy of HIV-associated neurodegenerative diseases such as dementia using reverse transcriptase inhibitors may thus be linked with expression of *ABCC*4 and *ABCC*5 transporters. In addition, a probenecid-sensitive efflux transport system for AZT, has been demonstrated in the BBB and this system may include the 2 probenecid-sensitive transporters (Takasawa et al. 1997). The localization of *ABCC*4 in the basolateral membrane of the choroid plexus epithelial cells and the apical membrane of brain capillary endothelial cells also adds weight to the argument that *ABCC*4 is perfectly poised to restrict brain penetration of topotecan and other substrates filtering from blood as well as preventing the accumulation of drug molecules in the cerebrospinal fluid (CSF) (Leggas et al. 2004). Endogenous anions such as taurocholate, dehydroepiandrosterone-3-sulfate, 2-estradiol-17β-glucuronide, and estrone sulfate as well as prostaglandins (which are produced by the choroids plexus) found in the brain are also *ABCC*4 substrates (Reid et al. 2003b; Zelcer et al. 2003).

Like the BBB and BCSFB, the placental membrane represents a structural barrier that protects the developing and sensitive fetus from distribution of certain xenobiotics in the maternal circulation. *ABCC*5 is shown to be expressed in human placenta by RT-PCR and Western blot analyses (Pascolo et al. 2003). Meyer et al., 2005 reported that

an increased amount of *ABCC*5 mRNA was detected in first and second trimester placentas in comparison with term placentas and MRP5 preferentially localizes in the basal membrane of syncytiotrophoblasts (Meyer Zu Schwabedissen et al. 2005; Ritter et al. 2005). A higher level of *ABCC*5 protein is also found in pre-term placentas (Meyer Zu Schwabedissen et al. 2005). The expression of this transporter in the membrane of syncytiotrophoblasts may indicate that the transporter may potentially regulate the intracellular concentration of the second messenger cGMP (and other cyclic nucleotides) and thereby partake in cGMP signaling. The extracellular level of cGMP is linked to cytotrophoblasts differentiation, control of fetoplacental vascular tone as well as NO-dependent placental angiogenesis. In addition, the presence of *ABCC*5 in placentas may influence nucleoside drug analogues and serve as a protective function against potential toxic agents.

As *ABCC*4 and *ABCC*5 may be involved in the efflux of some antiretroviral drugs aimed at targeting the disease pathway of human immunodeficiency virus (HIV), the temporal changes in expression levels of these 2 transporters in human macrophages may be due to HIV infection. In uninfected monocyte-derived macrophages (MDM), the expression levels of *ABCC*4 and *ABCC*5 were low and stable. In MDM cultures infected with HIV-1/BA-L, HIV infection in vitro causes a significant but transient increase in *ABCC*4 expression, which was also correlated with TNP-α production (Jorajuria et al. 2004). In contrast, *ABCC*5 and *ABCC*1 transcription levels were sharply increased with HIV replication not correlated with the production of cytokines for either TNF- α or IL-6 (Jorajuria et al. 2004).

2.4 Polymorphisms in coding regions of *ABCC*4 and *ABCC*5 gene loci

There are two main sources of variation data available: published reports in peer-reviewed journals and polymorphism/mutation databases. A few published reports have described polymorphisms in both the *ABCC*4 and *ABCC*5 gene loci (Table 5). Variations not strictly considered as SNPs like indels and repeats are nevertheless included in Table 5 and highlighted in grey. In 2002, Adachi et al. reviewed a selection of 10 *ABCC*4 variations and 5 *ABCC*5 variations by analyzing respective sequences in publicly available databases. One conservative substitution 1289 (Lys to Asn) reported by Lai et al., 2002 was not included. All these variants caused non-synonymous changes in amino acids (Adachi et al. 2002). These variants were however not verified as true SNPs occurring in human samples and no frequencies were reported. Nearly all reported variations could not be found in the public SNP databases. The variation I18L was likely to be the SNP e1 -49C>T. In the same year, Saito et al., screened 48 Japanese individuals in eight genes of the ATP-binding cassette, subfamily C (*ABCC*/MRP/CFTR). Among the 779 genetic polymorphisms, 230 are identified in *ABCC*4 and 85 in *ABCC*5. Within exons, a total of 11 *ABCC*4 SNPs and 7 *ABCC*5 variations (including 1 repeat sequence) are listed. Only two *ABCC*4 SNPs and none of the *ABCC*5 SNPs were nonsynonymous. No allele frequencies were reported. All the SNPs in this report were also listed in the NCBI dbSNP database.

More recently, Lockhart et al., 2003 reviewed polymorphisms in both nucleotide analogue transporters and listed a total of 10 *ABCC*4 and 10 *ABCC*5 variations (Lockhart et al. 2003). These variations were obtained by comparing various cDNA sequences of many sources in GenBank, including cell lines and across species e.g. mouse. Allele frequencies could not be obtained.

Only one report attempted to associate polymorphisms within the nucleotide analogue transporter genes with gene expression. Dazert et al., 2003 sequenced the *ABCC*5 gene in 21 Caucasian individuals and identified 20 polymorphisms (Dazert et al. 2003). Four exonic SNPs, all of which synonymous, were reported. They did not find any significant correlation between individual SNPs with *ABCC*5 mRNA expression, which could be a result of small sample size or the lack of power of association. Having a larger sample pool and using haplotypes instead of individual SNPs may have greater correlation power.

The evolution of polymorphism/mutation databases has matched the rapid growth of SNP data. From several independent SNP databases, we tabulated a list of coding SNPs in *ABCC*4 and *ABCC*5 (Table 6). Several variations reported by the publications listed in Table 5 were not found within these databases. A total of 46 variants was found in *ABCC*4, surpassing the number found from polymorphism/mutation databases while the number of variants found in *ABCC*5 is remarkably small, especially when the transcript lengths of these two proteins are similar. Moreover, the *ABCC*4 protein carries far more variants that cause a nonsynonymous change in amino acid. Like the majority of the published reports for both gene loci, the variations detected via alignment of sequencing data e.g. from overlapping clones may be artifacts or cannot be traced at a detectable frequency in a human population (Marth et al. 2001; Barnes 2002). Only 6 *ABCC*4 variations and 3 *ABCC*5 variations possess a minor allele frequency of more than 10%.

ABCC4 SNPs

Exon SNP Name	dbSNP rs#cluster id	Genotyped in our populations?	Exon	Function	Amino Acid Change	Adachi et al., 2002 (*Lai et al., 2002)	Saito et al., 2002	Lockhart et al., 2003	Dazert et al., 2003
e1 -49C>T	rs3751333	Yes	e1				e1 67C/T		
e1 52A>C	rs11568681		e1	nonsynonymous	Ile18Leu	I18L *			
e4 511G>T	rs4148460	Yes	e4	nonsynonymous	Gly171Cys		e4 205T/G	T513G	
e6 669C>T	rs899494	Yes	e6	synonymous	Ile223Ile		e6 48C/T	T669C	
e8 912T>G	rs2274407	Yes	e8	nonsynonymous	Asn304Lys		e8 1G/T	G906T	
e8 951G>A	rs2274406	Yes	e8	synonymous	Arg317Arg		e8 40G/A	G951T	
e8 969G>A	rs2274405	Yes	e8	synonymous	Ser323Ser		e8 58G/A	G969A	
e11 1497T>C	rs1557070	Yes	e11	synonymous	Tyr499Tyr			C1497T	
e15 Unknown			e15	nonsynonymous	Glu637Gly	E637G			
e15 Unknown			e15	nonsynonymous	Ser664Phe	S664F *			
e15 Unknown			e15	nonsynonymous	Gln677Arg	Q677R			
e16 Unknown			e16	nonsynonymous	Ser703Asn	S703N			
e18 2269A>G	rs3765534	Yes	e18	nonsynonymous	Lys757Glu		e18 56G/A	G2271A	
e18 Unknown			e18	nonsynonymous	Gly757Glu	G757E			
e22 Unknown			e21	nonsynonymous	Gly893Glu	G893E			
e22 2712G>A	rs1678339	Yes	e22	synonymous	Leu904Leu		e22 26A/G	A2712G	
e23 2844C>T	rs1189466	Yes	e23	synonymous	Phe948Phe		e23 38C/T	C2844T	
e23 Unknown			e23	nonsynonymous	Lys1103Arg	K1103R *			
e26 3348A>G	rs1751034	Yes	e26	synonymous	Lys1116Lys		e26 138A/G	A3348G	
e27 Unknown			e26	nonsynonymous	Asn1139Lys	N1139K *			
e30 Unknown			e30	nonsynonymous	Ser1267Gly	S1267G			
e30 Unknown			e30	nonsynonymous	Lys1289Arg	* 1289 (Lys to Arg)			
e31 4016G>T	rs3742106	Yes	e31				e31 146G/T		

ABCC5 SNPs

Exon SNP Name (* not true SNPs)	dbSNP rs#cluster id	Genotyped in our populations?	Exon	Function	Amino Acid Change	Adachi et al., 2002	Saito et al., 2002	Lockhart et al., 2003	Dazert et al., 2003
e3 Unknown			e3	nonsynonymous	Ala54Val	A54V			
e5 Unknown			e5	nonsynonymous	Pro176Arg	P176R			
* e5 GC527-8CG			e5	nonsynonymous	Arg176Pro			GC527-8CG	
e6 A723G			e6	synonymous	Ala241Ala			A723G	
e8 1145A>G		Yes	e8	synonymous	Gln382Gln			G1146A	e8 36156A/G
e9 1185C>T	rs1132776	Yes	e9	synonymous	Ala395Ala		e9 38C/T	C1185T	e9 39270C/T
* e9 AGC1198-200GGT			e9	nonsynonymous	Ser400Gly	S400G		AGC1198-200GGT	
e9 1200 C>T		Yes	e9	synonymous	Ser400Ser			C1200T	
e12 1743A>G		Yes	e12	nonsynonymous	Ile581Val	I584V		A1741G	e12 50138A/G
e12 1782C>T	rs939336	Yes	e12	synonymous	Cys594Cys		e12 21T/C	T1782C	
e25 3606 C>A		Yes	e25	nonsynonymous	Tyr1202*				
e25 3624C>T	rs3749442	Yes	e25	synonymous	Leu1208Leu		e25 120/T	T3624C	e25 75087T/C
e29 4148 C>A		Yes	e29	nonsynonymous	Thr1383Asp	N1383T		C4148A	
e30 4896G>A	rs3749445	Yes	e30				e30 684G/A		
e30 5159 C>T		Yes	e30				e30 947C/T		
* e30 5357-5362 (TC)6-8			e30				e30 (1145-1160) (TC)6-8		
e30 5557A>G	rs562	Yes	e30				e30 1345A/G		

Table 5. Coding polymorphisms/mutations in *ABCC*4 and *ABCC*5 as reported from published reports. Adachi et al., 2002 reviewed *ABCC*4 sequences in 5 studies including Schuetz et al., 1999, Kruh et al., 2003, Lai et al., 2002, and van Aubel 2002 for SNPs (Schuetz et al. 1999; Adachi et al. 2002; Lai and Tan 2002; van Aubel et al. 2002; Kruh and Belinsky 2003), and *ABCC*5 sequences in another 5 studies from Suzuki et al., 1997, Belinsky et al., 1998, Wijnholds et al. 2000, McAleer et al. 1999, Jedlitchsky et al., 2000 (Belinsky et al. 1998; McAleer et al. 1999; Jedlitschky et al. 2000; Suzuki et al. 2000; Wijnholds et al. 2000). By chance the polymorphism Lys1289Arg recorded by Lai et al., 2002 was not listed in Adachi's review (marked by *). Lockhart, on the other hand reviewed *ABCC*4 polymorphisms in 2 studies from Saito et al., 2002 and Lee et al., 1998 (Lee et al. 1998; Saito et al. 2002a). Lockhart also reviewed *ABCC*5 polymorphisms from Belinsky et al., 1998, McAleer et al., 1999, Wijnholds et al. 2000, and Suzuki et al., 2000 (Belinsky et al. 1998; McAleer et al. 1999; Suzuki et al. 2000; Wijnholds et al. 2000). However, he did not include *ABCC*5 SNPs recorded by Saito et al., 2002.

More work is needed to characterize both *ABCC*4 and *ABCC*5 polymorphisms. There are certain concerns on the reliability of SNP information in public databases as well as comparative sequence data. Many of the variations reported as 'SNPs' in review publications and polymorphism/mutation databases were generated from comparative data of DNA sequences from overlapping clones and may therefore be mutations in cell lines used or sequencing errors, especially in small samples. The 'false-positive' variations may not be found in any human population at a detectable frequency (Marth et al. 2001; Barnes 2002). Re-genotyping may be required to verify these reported polymorphisms as true SNPs occurring in normal human populations and the International Hapmap Project is one major sequencing effort to yield data of verified SNPs. It remains elusive as to which of these *ABCC*4 and *ABCC*5 polymorphisms may have an impact on drug disposition or pharmacodynamics in vivo. In candidate gene-based studies such as this one, there are many more SNPs than can be investigated for every gene. A SNP selection strategy must be in place to permit a feasible list of SNPs to be investigated. Sources to look for variants include published reports and SNP databases. Sequencing pooled samples may also be a good complementary strategy. Validated SNPs are to be selected as well as common SNPs

that are known to exist at a frequency of more than 5-10% in the general population. SNPs that reside in exonic regions of the candidate genes especially those polymorphisms leading to a non-synonymous change in amino acid have a higher probability of association with a disease or drug response phenotype. SNPs residing at exon-intron junctions or at intron sites important for RNA stability (such as the 5'- and 3'- untranslated regions) may also be considered. Lastly SNPs within other intronic regions may be selected to achieve a high resolution sufficient for linkage disequilibrium analysis.

ABCC4 SNPs

	Exon SNP Name	dbSNP rs#	reference JSNP id	JSNP id	TSC SNP id	Genotyped in our studies?	Location	Function	Allele 1 - encoded amino acid	Allele 2 - encoded amino acid	Codon position	Amino Acid Change	Heterozygosity	Reported variation data in dbSNP		
														Number of Chromosomes	Major Allele Freq	Minor Allele Freq
1	e1 52A>C	rs11568681	-	-	-	-	exon 1	nonsynonymous	A - Ile [I]	C - Leu [L]	1	Ile18Leu	0.06	552	C - 0.969	A - 0.031
2	e3 232G>C	rs11568689	-	-	-	-	exon 3	nonsynonymous	G - Ala [A]	C - Pro [P]	1	Ala78Pro	0.004	552	C - 0.998	G - 0.02
3	e4 511G>T	rs4148460	ssj0000492	-	-	Yes	exon 4	nonsynonymous	G - Gly [G]	T - Cys [C]	1	Gly171Cys	N.D.	96	-	-
4	e5 551C>T	rs11568657	-	-	-	-	exon 5	nonsynonymous	C - Thr [T]	T - Met [M]	2	Thr184Met	0.007	552	T - 0.996	C - 0.004
5	e5 559T>G	rs11568658	-	-	-	-	exon 5	nonsynonymous	T - Trp [W]	G - Gly [G]	1	Trp187Gly	0.103	552	G - 0.946	T - 0.054
6	e6 669C>T	rs899494	ssj0000521	IMS-JST179043	TSC0174780	Yes	exon 6	synonymous	C - Ile [I]	T - Ile [I]	3	Ile223Ile	0.313	712	G - 0.806	A - 0.194
7	e6 717C>T	rs11568674	-	-	-	-	exon 6	synonymous	C - Ala [A]	T - Ala [A]	3	Ala239Ala	0.004	552	T - 0.998	C - 0.002
8	e6 732G>A	rs11568679	-	-	-	-	exon 6	synonymous	G - Leu [L]	A - Leu [L]	3	Leu244Leu	0.004	552	A - 0.998	G - 0.002
9	e7 877G>A	rs11568684	-	-	-	-	exon 7	nonsynonymous	G - Glu [E]	A - Lys [K]	1	Glu293Lys	0.004	552	A - 0.998	G - 0.002
10	e8 912T>G	rs2274407	ssj0000529	IMS-JST070256	-	Yes	exon 8	nonsynonymous	T - Asn [N]	G - Lys [K]	3	Asn304Lys	0.174	900	C - 0.904	A - 0.096
11	e8 951G>A	rs2274406	ssj0000530	IMS-JST070255	-	Yes	exon 8	synonymous	G - Arg [R]	A - Arg [R]	3	Arg317Arg	0.499	900	T - 0.522	C - 0.478
12	e8 969G>A	rs2274405	ssj0000531	IMS-JST070254	-	Yes	exon 8	synonymous	G - Ser [S]	A - Ser [S]	3	Ser323Ser	0.49	902	C - 0.572	T - 0.428
13	e8 1035A>G	rs11568703	-	-	-	-	exon 8	synonymous	A - Val [V]	G - Val [V]	3	Val345Val	0.007	552	G - 0.996	A - 0.004
14	e8 1067T>C	rs11568701	-	-	-	-	exon 8	nonsynonymous	T - Met [M]	C - Thr [T]	2	Met356Thr	0.004	552	C - 0.998	T - 0.002
15	e9 1208T>C	rs11568705	-	-	-	-	exon 9	nonsynonymous	T - Leu [L]	C - Pro [P]	2	Leu403Pro	0.004	552	C - 0.998	T - 0.002
16	e11 1458A>G	rs11568670	-	-	-	-	exon 11	synonymous	A - Ser [S]	G - Ser [S]	3	Ser486Ser	0.004	552	G - 0.998	A - 0.002
17	e11 1460A>G	rs11568668	-	-	-	-	exon 11	nonsynonymous	A - Glu [E]	G - Gly [G]	2	Glu487Gly	0.004	552	G - 0.998	A - 0.002
18	e11 1492G>A	rs11568669	-	-	-	-	exon 11	nonsynonymous	G - Glu [E]	A - Lys [K]	1	Glu498Lys	0.014	552	A - 0.993	G - 0.007
19	e11 1497T>C	rs1557070	-	-	TSC0455420	Yes	exon 11	synonymous	T - Tyr [Y]	C - Tyr [Y]	3	Tyr499Tyr	0.111	564	C - 0.941	T - 0.059
20	e14 1737C>T	rs11568664	-	-	-	-	exon 14	synonymous	C - Cys [C]	T - Cys [C]	3	Cys579Cys	0.004	552	T - 0.998	C - 0.002
21	e15 1875G>A	rs11568699	-	-	-	-	exon 15	nonsynonymous	G - Met [M]	A - Ile [I]	3	Met625Ile	0.004	552	A - 0.998	G - 0.002
22	e15 2000T>C	rs11568697	-	-	-	-	exon 15	nonsynonymous	T - Leu [L]	C - Pro [P]	2	Leu667Pro	0.004	552	C - 0.998	T - 0.002
23	e15 2001T>C	rs11568698	-	-	-	-	exon 15	synonymous	T - Pro [P]	C - Pro [P]	3	Pro667Pro	0.007	552	C - 0.996	T - 0.004
24	e16 2100T>C	rs11568666	-	-	-	-	exon 16	synonymous	T - Ala [A]	C - Ala [A]	3	Ala700Ala	0.007	552	C - 0.996	T - 0.004
25	e18 2230G>A	rs11568646	-	-	-	-	exon 18	nonsynonymous	G - Val [V]	A - Met [M]	1	Val744Met	0.029	552	A - 0.986	G - 0.014
26	e18 2269A>G	rs3765534	ssj0000585	IMS-JST119723	-	Yes	exon 18	nonsynonymous	A - Lys [K]	G - Glu [E]	1	Lys757Glu	0.02	900	C - 0.99	T - 0.01
27	e19 2364T>C	rs11568709	-	-	-	-	exon 19	synonymous	T - Tyr [Y]	C - Tyr [Y]	3	Tyr788Tyr	0.004	552	C - 0.998	T - 0.002
28	e20 2459T>G	rs11568659	-	-	-	-	exon 20	nonsynonymous	T - Ile [I]	G - Arg [R]	2	Ile820Arg	0.004	552	G - 0.998	T - 0.002
29	e21 2560T>G	rs11568694	-	-	-	-	exon 21	nonsynonymous	T - Phe [F]	G - Val [V]	1	Phe854Val	0.004	552	G - 0.998	T - 0.002
30	e21 2577T>C	rs11568691	-	-	-	-	exon 21	synonymous	T - Ala [A]	C - Ala [A]	3	Ala859Ala	0.004	552	C - 0.998	T - 0.002
31	e21 2578A>G	rs11568692	-	-	-	-	exon 21	nonsynonymous	A - Met [M]	G - Val [V]	1	Met860Val	0.007	552	G - 0.996	A - 0.004
32	e22 2698T>G	rs11568673	-	-	-	-	exon 22	nonsynonymous	T - Leu [L]	G - Val [V]	1	Leu900Val	0.007	552	G - 0.996	T - 0.004
33	e22 2712G>A	rs1678339	ssj0000655	IMS-JST070250	-	Yes	exon 22	synonymous	G - Leu [L]	A - Leu [L]	3	Leu904Leu	0.338	705	C - 0.784	T - 0.216

Table 6. Coding SNPs found from public polymorphism/mutation databases. The following polymorphism data has been collectively tabulated from these sources: The Single Nucleotide Polymorphism database (dbSNP) [http://www.ncbi.nlm.nih.gov/projects/SNP/]; The SNP Consortium (TSC) [http://snp.cshl.org/]; Japanese SNP (JSNP database [http://snp.ims.u-tokyo.ac.jp/]

ABCC4 SNPs

	Exon SNP Name	dbSNP rs#	reference JSNP id	JSNP id	TSC SNP id	Genotyped in our studies?	Location	Function	Allele 1 - encoded amino acid	Allele 2 - encoded amino acid	Codon position	Amino Acid Change	Heterozygosity	Reported variation data in dbSNP		
														Number of Chromosomes	Major Allele Freq	Minor Allele Freq
34	e23 2844C>T	rs1189466	ssj0000657	IMS-JST070249	-	Yes	exon 23	synonymous	C - Phe [F]	T - Phe [F]	3	Phe948Phe	0.332	711	G - 0.789	A - 0.211
35	e23 2847T>C	rs11568708	-	-	-	-	exon 23	synonymous	T - Ala [A]	C - Ala [A]	3	Ala949Ala	0.011	552	C - 0.995	T - 0.005
36	e23 2867C>G	rs11568707	-	-	-	-	exon 23	nonsynonymous	C - Ser [S]	G - Cys [C]	2	Ser956Cys	0.004	552	G - 0.998	C - 0.002
37	e26 3211A>G	rs11568653	-	-	-	-	exon 26	nonsynonymous	A - Ile [I]	G - Val [V]	1	Ile1071Val	0.004	552	G - 0.998	A - 0.002
38	e26 3255A>C	rs11568652	-	-	-	-	exon 26	synonymous	A - Ile [I]	C - Ile [I]	3	Ile1085Ile	0.011	552	C - 0.995	A - 0.005
39	e26 3310C>T	rs11568655	-	-	-	-	exon 26	synonymous	C - Leu [L]	T - Leu [L]	1	Leu1104Leu	0.06	552	T - 0.969	C - 0.031
40	e26 3348A>G	rs1751034	ssj0000672	-	-	Yes	exon 26	synonymous	A - Lys [K]	G - Lys [K]	3	Lys1116Lys	0.336	654	T - 0.786	C - 0.214
41	e27 3425T>C	rs11568644	-	-	-	-	exon 27	nonsynonymous	T - Met [M]	C - Thr [T]	2	Met1142Thr	0.004	552	C - 0.998	T - 0.002
42	e28 3609A>G	rs11568695	-	-	-	-	exon 28	synonymous	A - Ala [A]	G - Ala [A]	3	Ala1203Ala	0.116	552	G - 0.938	A - 0.062
43	e29 3659A>G	rs11568639	-	-	-	-	exon 29	nonsynonymous	A - Gln [Q]	G - Arg [R]	2	Gln1220Arg	0.004	552	G - 0.998	A - 0.002
44	e29 3723T>C	rs11568640	-	-	-	-	exon 29	synonymous	T - Ser [S]	C - Ser [S]	3	Ser1241Ser	0.004	552	C - 0.998	T - 0.002
45	e30 3774A>G	rs11568704	-	-	-	-	exon 30	synonymous	A - Pro [P]	G - Pro [P]	3	Pro1258Pro	0.022	552	G - 0.989	A - 0.011
46	e31 3941G>A	rs11568688	-	-	-	-	exon 31	nonsynonymous	G - Arg [R]	A - Gln [Q]	2	Arg1314Gln	0.004	552	A - 0.998	G - 0.002
47	e31 4016G>T	rs3742106	ssj0000696	IMS-JST093204	-	Yes	exon 31									

ABCC5 SNPs

	Exon SNP Name	dbSNP rs#	reference JSNP id	JSNP id	TSC SNP id	Genotyped in our studies?	Location	Function	Allele 1 - encoded amino acid	Allele 2 - encoded amino acid	Codon position	Amino Acid Change	Heterozygosity	Reported variation data in dbSNP		
														Number of Chromosomes	Major Allele Freq	Minor Allele Freq
1	e7 855C>G	rs11708427	-	-	-	-	exon 7	nonsynonymous	C - His [H]	G - Gln [Q]	3	His285Gln	N.D.	2	-	-
2	e8 1145A>G	rs7636910	-	-	-	Yes	exon 8	synonymous	A - Gln [Q]	G - Gln [Q]	3	Gln382Gln	N.D.	2	-	-
3	e9 1185C>T	rs1132776	ssj0000733	IMS-JST066967	-	Yes	exon 9	synonymous	T - Ala [A]	C - Ala [A]	3	Ala395Ala	0.38	200	C - 0.746	T - 0.254
4	e9 1200C>T	rs1053386	-	-	-	Yes	exon 9	synonymous	C - Ser [S]	T - Ser [S]	3	Ser400Ser	N.D.	4	-	-
5	e12 1782T>C	rs939336	ssj0000744	IMS-JST066966	TSC0359353	Yes	exon 12	synonymous	T - Cys [C]	C - Cys [C]	3	Cys594Cys	0.36	1662	C - 0.768	T - 0.232
6	e 23 3303C>T	rs11552530	-	-	-	-	exon 23	nonsynonymous	C - Trp [W]	T - Arg [R]	1	Trp1101Arg	N.D.	2	-	-
7	e25 3606C>A	rs1053351	-	-	-	Yes	exon 25	nonsynonymous	C - *	A - Tyr [Y]	3	Tyr1202*	N.D.	9	-	-
8	e25 3624C>T	rs3749442	ssj0000769	IMS-JST101281		Yes	exon 25	synonymous	C - Leu [L]	T - Leu [L]	3	Leu1208Leu	0.37	1446	C - 0.76	T - 0.24
9	e29 4148C>A	rs1053387	-	-	-	Yes	exon 29	nonsynonymous	C - Thr [T]	A - Asn [N]	2	Thr1383Asn	N.D.	184	C - 1	A - 0

Table 6 (continued).

2.5 Clinical implications of *ABCC*4 and *ABCC*5 in drug efflux

Based on substrate profiles of *ABCC*4 and *ABCC*5, clinical drugs with structural similarity to those used in experiments may also be potential substrates for *ABCC*4/*ABCC*5. These clinical drugs are mostly antitumour and antiviral agents. In association studies linking genotype to clinical drug response or disease, this selection of drugs in current clinical use may serve to be useful.

1) Nucleotide and nucleoside analogues

Nucleoside Reverse Transcriptase Inhibitors

Treatment of nucleoside reverse transcriptase inhibitors (NRTIs) has been the cornerstone of HIV therapy since the first NRTI was introduced in 1987. With the availability of protease inhibitors (PIs) in 1997, NRTIs became the foundation of highly active antiretroviral therapy (HAART). With at least 7 NRTIs and NtRTIs on the market today, selection of these agents requires the careful and thoughtful consideration of efficacy, safety, resistance, and long-term tolerance profiles. NRTI/NtRTIs are structurally similar to the purine nucleosides adenosine and guanine, and the pyrimidine nucleosides thymidine and cytidine. They, however, lack the terminal 3'-OH group of the nucleoside. The mechanism of action of these NRTI/NtRTIs is by competitive inhibition. When these analogues are phosphorylated intracellularly and incorporated into the growing viral DNA chain, synthesis of DNA by reverse transcriptase (RT) is terminated. Viral RT mediates replication of viral DNA from viral RNA. Thus viral replication in the infected cell can no longer proceed.

Hepatitis B antiviral agents

Nucleoside/nucleotide analogues suppress HBV replication through inhibition of HBV DNA polymerase. A dideoxynucleoside reverse transcriptase inhibitor, lamivudine inhibits viral DNA replication. It is converted intracellularly to its active metabolite, b-L-3TC-triphosphate. Emerging treatment options include Adefovir, Dipivoxil and Entecavir. The adenine nucleotide analogue adefovir dipivoxil and the guanosine nucleoside analogue entecavir are oral antiviral drugs with activity against both wild-type and lamivudine-resistant HBV. Other experimental nucleoside analogue drugs include Telbivudine (LdT; Idenix; Phase III), Clevudine (L-FMAU; Gilead; Phase II), Elvucitabine (ACH 126,443; Acillon; Phase II), Valtorcitabine (Idenix; Phase II), Amdoxovir (DAPD, Triangle; Phase II), acivir (Pharmasset; Phase II), MCC478 (Elli Lilly; Phase I), MIV 210 (Medivir; Phase I), and Hepavir B (Ribapharm; Phase I).

Hepatitis C antiviral agents

Ribavirin [1-β-D-ribofuranosyl-1H-1,2,4-triazole-3-carboxamide] is a nucleoside analogue with antiviral activity used in combination therapy (primarily with interferon) for the treatment of hepatitis C. The U.S. Food and Drug Administration (FDA) in 1998 approved Rebetron (combination of interferon alfa-2band ribavirin) for the treatment of individuals with chronic hepatitis C [http://www.fda.gov/cder/foi/label/2002/20903slr025_Rebetron_lbl.pdf] and on 24th Feb 2004 gave an approval letter for Ribaspheret (Ribavirin as monotherapy).

Antitumour Cytosine Analogues

Gemcitabine (2',2'-difluorudeoxycytidine, dFdC) is an S-phase nucleoside cytidine analogue that is currently considered to be the single agent of choice in advanced pancreatic cancer and is being investigated in numerous trials as part of novel combination regimens [http://www.fda.gov/cder/foi/label/1998/20509lbl.pdf]. Gemcitabine has three mechanisms of action: it competes for incorporation into DNA, thereby inhibiting the synthesis of DNA; it prevents DNA repair by masked termination; and it undergoes self-potentiation. Gemcitabine primarily and irreversibly inhibits ribonucleotide reductase, the enzyme responsible for catalyzing the biosynthesis of deoxyribose nucleotides, which are essential for DNA replication. A drug recently abandoned in Phase II clinical trials is Tezacitabine [(E)-2'-deoxy-2'-(fluoromethylene) cytidine (FMdC)] (designated orphan drug in 2005 by FDA), which acts similarly to gemcitabine. Also within this class is Troxacitabine (Troxatyl) (designated orphan drug in 2005 by FDA), a dioxolane nucleoside analogue of cytidine that is incorporated into DNA during replication, inhibiting DNA polymerase and DNA synthesis. Like gemcitabine it is intracellularly phosphorylated by deoxycytidine kinase to its active metabolite troxacitabine triphosphate. Unlike gemcitabine, however, it is not incorporated into RNA, it does not inhibit ribonucleotide reductase and it does not require nucleoside transporters, but enters cells by passive diffusion. Another pyrimidine nucleoside analogue of cytidine, Azacitidine (Vidaza, Pharmion Corporation, Boulder, CO) [4-amino-1-ß-D-ribofuranosyl-s-triazin-2(1H)-one] is recently marketed in 2004 as an orphan drug for the use in myelodysplastic syndromes subtypes (MDS) [http://www.fda.gov/cder/foi/label/2004/050794lbl.pdf]. Known as the first of a new class of drugs known as DNA "hypomethylating" or "demethylating" agents, it is

thought to act by inhibiting DNA methyltransferase, the enzyme that is responsible for methylating newly synthesized DNA. Since methylation of DNA is a major mechanism regulating gene expression and increased DNA methylation can result in silencing of tumor suppressor genes, inhibition of methytransferase ("demethylation") by azatidine may therefore restore the expression of genes silenced by methylation (Issa et al. 2005; Kaminskas et al. 2005a; Kaminskas et al. 2005b).

2) Non-nucleotide and non-nucleoside analogues

Nonsteroidal Anti-inflammatory Drugs (NSAIDs)

Cellular efflux and vesicular uptake studies showed that both *ABCC*4 and *ABCC*5 are inhibited by physiological relevant concentrations of nonsteroidal anti-inflammatory drugs (NSAIDs) dipyramidole and indomethacin (Reid et al. 2003a; Reid et al. 2003b; Jedlitschky et al. 2004) (albeit by different degrees). Presence of *ABCC*4 and *ABCC*5 in platelet membrane vesicles and granules as well as (porcine) coronary and pulmonary arteries also indicate a role of these two transporters in treatment of cardiovascular diseases including anticoagulant therapy.

Uricosuric agents

Reid et al., 2003 showed that two uricosuric agents, probenecid, and to a lesser extent sulfinpyrazone inhibit *ABCC*5 more than *ABCC*4. Both transporters are located in the kidney with *ABCC*4 protein expressed at the apical (brush border) membrane of the human renal proximal tubules. Uricosuric agents lower uric acid levels in patients with hyperuricemia and gout. Long term usage of uricosuric agents lowers uric acid levels and prevents the formation of uric acid crystals

PDE5 Inhibitors

In 1988, FDA approved sildenafil citrate, a selective PDE5 inhibitor for the treatment of erectile dysfunction. By inhibiting the degradation of cGMP, PDE5 inhibitors prolong the activity of this cyclic nucleotide second messenger within the cavernous vasculature and smooth musculature, thus potentiating the erectile response. Two other drugs in this class are valdenafil and tadalafil.

Leukotrienes Receptor Antagonists (LTRAs)

The selective and competitive antagonist of leukotriene D_4, MK571 is also a strong inhibitor of *ABCC*4- (Chen et al. 2002; Reid et al. 2003a), but not *ABCC*5-mediated cGMP transport (Jedlitschky et al. 2000; Klokouzas et al. 2003). There are three commercially available but structurally distinct drugs: montelukast and zafirlukast and pranlukast, which is available only in Asia. The LTRA drugs also known as leukotriene modifiers demonstrate both bronchodilator and anti-inflammatory properties and are used in the management of mild, persistent asthma.

Topoisomerase I inhibitors

Most inhibitors of type I topoisomerases are a new class of antineoplastic agents derived from campothecins. Camptothecins are alkaloids extracted from plants like Camptotheca acuminata. Drugs which are in clinical studies include natural/semisynthetic/synthetic derivatives of campthothecin: irinotecan [Camptosar, CPT-11]; topotecan [Hycamptin, TPT, 9-dimethyl-amino-methyl-10-hydroxycamptothecin, 7-ethyl-10-[4-(1-piperidino)-1-piperidino]carbonyloxycamptothecin], lurtotecan [NX211, OSI-211], exatecan [Ensiv, DX-8951f] and rubitecan [Orathecin, 9-nitrocamptothecin, 9NC, RFS 2000]. DNA

topoisomerases (types I and II) are enzymes in the cell nucleus that control DNA topology during nuclear processes such as replication, recombination, and repair. Type I DNA topoisomerase creates transient (reversible) single-stranded breaks in double-stranded DNA to relieve torsional strain inherent in supercoiled DNA double helix and binds covalently with the 3'-DNA terminus of the cleaved DNA strand, forming a complex. This complex can be cleaved and topoisomerase enzyme reanneals the broken DNA strands for transcription to proceed. Type I topoisomerase inhibitors are proposed to interact with the DNA-DNA topoisomerase cleavable complex during the S-phase of DNA synthesis. This interaction prevents the topoisomerase enzyme from religating the single-strand breaks. The accumulation of the ternary complex of DNA, topoisomerase type I inhibitor and DNA topoisomerase results in irreversible defects in DNA replication and subsequent cell cycle arrest and cell death. It was found that *ABCC*4, but not *ABCC*5 confers resistance to the topoisomerase inhibitor irinotecan and its active metabolite SN-38 *in vitro* (Norris et al. 2005). Remarkably the *ABCC*4-mediated resistance to irinotecan and SN-38 was relatively specific with less notable resistance to topotecan and camptothecin, even though all these drugs belong to the same class.

The involvement of *ABCC*4 and *ABCC*5 transporters in drug resistance in clinical practice remains speculative but is supported by their substrate profiles and expression in specific tissues. Transport assays with recombinant *ABCC*4 or *ABCC*5 proteins may provide further evidence for the roles of *ABCC*4 and *ABCC*5 as transporters for the numerous drugs listed above. The identification of any polymorphisms within the *ABCC*4 or *ABCC*5 affecting any function may reveal the roles of both transporters as well as uncover new mechanisms of action of these drugs. Contrasting reports about substrate profiles of *ABCC*4 and *ABCC*5 have recently surfaced. In contrast to at least

two reports of both *ABCC*4 and *ABCC*5's ability to transport cyclic nucleotides such cGMP with high affinity (Jedlitschky et al. 2000; Chen et al. 2001), Reid's vesicular transport experiments showed low uptake of cGMP in *ABCC*4- or *ABCC*5-overexpressing cells (Reid et al. 2003a). In addition, Reid also found that *ABCC*5 is insensitive to 3 inhibitors of cGMP-specific phosphodiesterase 5 (PDE5) (zaprinast, sildenafil, and trequinsin), while Jedlitschky reported a significant inhibition by the same substrates. These discrepancies could possibly be due to the use of different in vitro systems. The polymorphisms within the *ABCC*4 and *ABCC*5 cDNA independently isolated in different labs may also be another reason for these conflicting reports. An in-depth study of *ABCC*4 and *ABCC*5 polymorphisms is therefore needed to shed some light on these apparent discrepancies.

Chapter 3 Background and specific aims of studies

3.1 Workflow of studies

To facilitate future association studies of the 2 nucleotide analogue transporters, *ABCC*4 and *ABCC*5 for future gene-based association studies, a step-by-step approach is devised from the development of a cost-effective SNP genotyping method to analyze multiple SNPs to characterizing the haplotype and linkage disequilibrium profiles of the 2 gene loci to finally achieving a manageable set of population-specific tagging SNPs and inferring signatures of positive selection in polymorphisms (Figure 5).

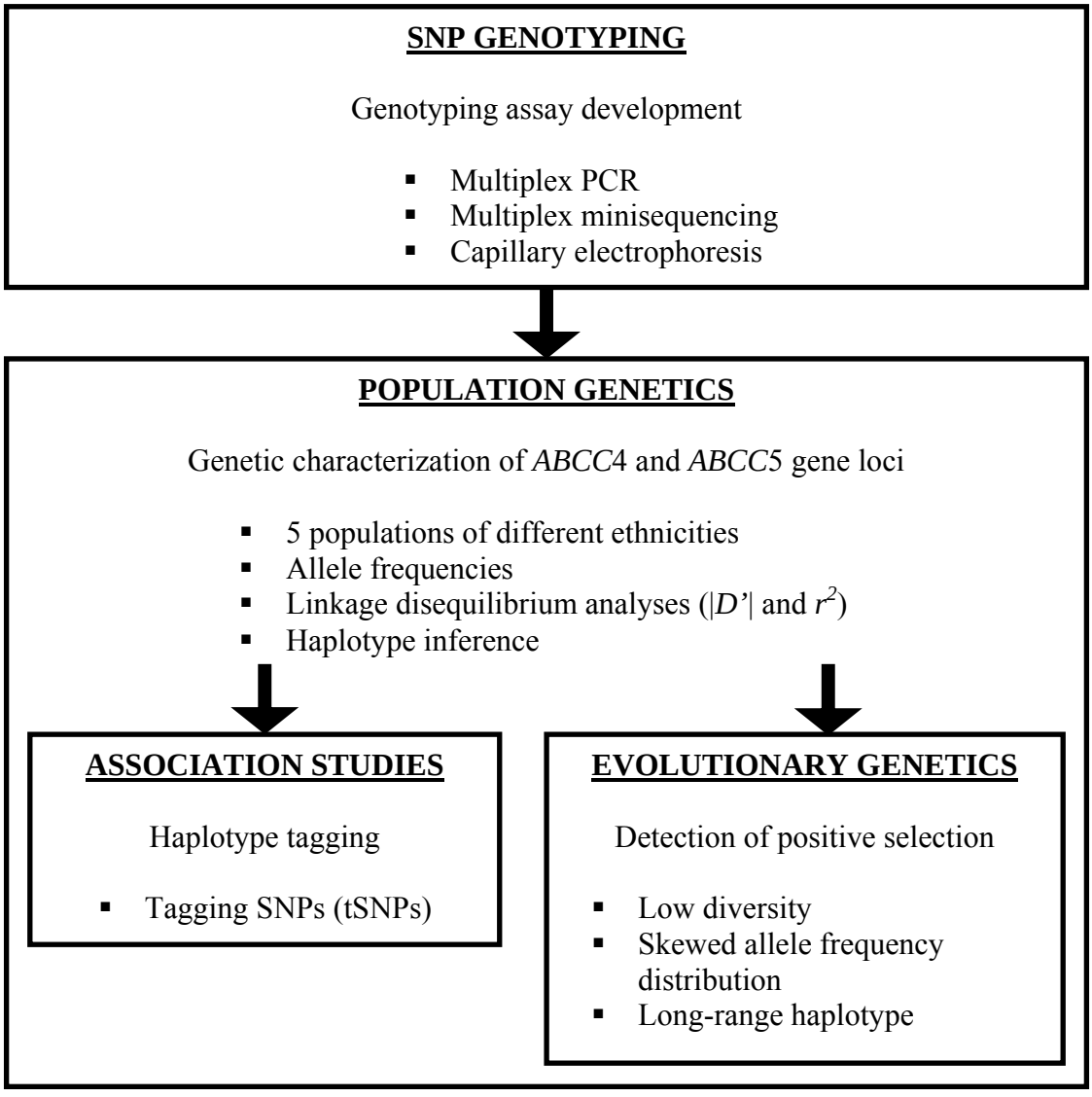


Figure 5. Overview of studies.

3.2 Genotyping assay development

Responses to different drugs can vary widely among different individuals as a result of genetic variations in drug-metabolizing enzymes, transporters, receptors, and/or other cofactors. The multidrug resistance 1 (MDR1) transporter, a well-characterized member of the ATP-binding cassette superfamily, was shown to efflux a wide variety of structurally and functionally unrelated drugs, including anticancer, antiarrhythmic, antidepressant, antipsychotic, and antiviral agents. The pharmacogenetics of the MDR1 multidrug transporter have recently received much scientific attention. Several single nucleotide polymorphisms (SNPs) have been identified in the MDR1 gene; some occur only in specific ethnic groups, whereas others occur in all ethnic groups but at significantly different allele frequencies among the different races [see Ref. (Tang et al. 2002) and references therein]. Nonetheless, the functional significance of these SNPs remains unclear. Various functional associations, some paradoxical, have been observed between the synonymous SNP (exon 26 3435C>T) and MDR1 protein expression and plasma drug concentrations (Hoffmeyer et al. 2000; Hitzl et al. 2001; Kim et al. 2001; Sakaeda et al. 2001), drug-induced side effects (Roberts et al. 2002), and drug response (Fellay et al. 2002). The SNP exon 26 3435T allele has been associated with lower MDR1 expression in the duodenum (Hoffmeyer et al. 2000), leukocytes (Hitzl et al. 2001), and placental tissues (Tanabe et al. 2001), leading to lower rhodamine efflux (Hitzl et al. 2001) and increased plasma digoxin concentrations (Hoffmeyer et al. 2000). In addition, early-onset Parkinson patients have higher frequency of the SNP exon 26 3435T allele compared with late-onset patients or unaffected controls (Furuno et al. 2002). However, although this same allele has been associated with lower MDR1 expression in peripheral blood mononuclear cells and better response to HIV-1 drugs, it has also been associated

with lower plasma concentrations of nelfinavir (Fellay et al. 2002). Additionally, the SNP exon 26 3435T allele has been associated with an increased risk of nortriptyline-induced postural hypertension, although blood concentrations of nortriptyline in these individuals were not significantly different from those in individuals carrying the C allele (Roberts et al. 2002). Furthermore, no association has been demonstrated between the SNP exon 26 3435C>T polymorphism and cyclosporin A efficacy in renal transplant patients (von Ahsen et al. 2001) or cyclosporin A pharmacokinetics (Min and Ellingrod 2002) in 14 healthy individuals. Together, these studies suggest that SNP exon 26 3435C>T itself may not be the causal variant producing these observed functional differences.

Analyses of other SNPs within the MDR1 gene have revealed haplotype frequency differences among populations (Kim et al. 2001; Tang et al. 2002). The linkage disequilibrium (LD) spanning the 40 kb between SNPs exon 12 1236T>C and exon 26 3435C>T was also found to differ among the Chinese, Malays, and Indians (Tang et al. 2002). Variations in LD blocks among different ethnic groups and the existence of ethnic-specific SNPs suggest that the current confusing association of SNP exon 26 3435C>T with different functional changes may be attributable to strong LD between SNP exon 26 3435C>T and different, as yet unidentified, causal SNPs within the different LD blocks in the different study populations. Analyses of MDR1 haplotypes rather than genotypes may provide additional insights in determining associations with functional differences and may assist in discriminating between surrogate SNPs and causative variants. It would be useful to determine the haplotype structure of the entire 100-kb MDR1 gene in the different ethnic populations and to study the relationship between MDR1 haplotypes and drug response.

The primary aim is to first develop a rapid and robust assay to simultaneously genotype seven SNPs across the MDR1 gene. The seven SNPs span 100 kb of the gene and are located in five genomic regions that are potentially functional (promoter and exons 12, 21, 26, and 28). In this study, the principle of the genotyping assay is based on minisequencing or single-base extension (SBE) due to previous reports of its high accuracy (Syvanen 2001). Besides the ability to genotype multiple SNPs in a single reaction, the assay has to be scalable to 96 samples, sensitive to detect polymorphisms in small quantities of DNA and cost-effective. By coupling to an automated capillary electrophoretic platform, allelic discrimination of the different sizes of minisequencing products can be fast and efficient. Although this assay is first established using MDR1 as a candidate gene, it will be quickly modified and be the driving force for the haplotype and linkage disequilibrium studies of *ABCC*4 and *ABCC*5 gene loci.

3.3 Genetic characterization of *ABCC*4 and *ABCC*5 gene loci

More than one million SNPs have been validated in four human populations from the efforts of the International Hapmap Project [http://www.hapmap.org/] (2003). Such construction of high-density SNP maps has facilitated the testing of genetic variants of a potentially contributing gene in an association study through candidate gene studies. The candidate gene approach allows LD analyses and makes use of patterns of LD inherent in a gene region of interest to select a subset of tagging SNPs (tSNPs) to comprehensively describe common variations within the gene (Goldstein et al. 2003). This significantly reduces genotyping effort with only a moderate loss of power (Johnson et al. 2001; Zhang et al. 2002). It has been estimated that at least 400,000 tSNPs would still be required genome-wide (Carlson et al. 2004), and

genotyping this number of SNPs would still pose a major challenge in looking for causal gene loci. Recently the inference of natural selection as a strategy to look for functionally important variants has been of interest (Bamshad and Wooding 2003; Clark 2003). Under the influence of positive selective forces, a variant that increases the fitness of an individual to its environment undergoes fixation and leaves behind evolutionary footprints of low diversity and skewed allele frequency distribution (Altshuler and Clark 2005). These footprints are therefore trademarks of genes or loci that have undergone evolution under positive selection in recent history of humans due to adaptation of early humans as they dispersed out of Africa to specific geographical environments. Genes or loci that might underlie variation in disease resistance or drug transport can thus be identified through these changes (Bamshad and Wooding 2003) and there is increasing evidence of recent positive selective events on ABC drug transporters (Tang et al. 2004; Wang et al. 2005).

Genetic variations in the MRP genes have been implicated in differences in drug response between individuals, and mutations in several genes have been identified in several human disorders, such as cystic fibrosis, sterol and bile salt transport deficiencies and retinal degeneration (Dean et al. 2001). Hence, this study hypothesizes that single nucleotide polymorphisms (SNPs) within the *ABCC*4 or *ABCC*5 gene may potentially affect its expression and/or alter protein function/structure, leading to changes in the transport profiles and response to specific drugs and compounds. At present, very little information is available underlying the *ABCC*4 and *ABCC*5 transporter polymorphisms resulting in inter-individual differences in drug response although it has been shown that there is large variability in *ABCC*4 expression in pediatric leukemias (Sampath et al. 2002) and a single study

of 21 Caucasian individuals found no statistically significant correlation between individual SNPs and mRNA expression of the *ABCC*5 gene (Dazert et al. 2003)..

There is increasing interest in utilizing linkage disequilibrium (LD) and haplotype analyses for detecting associations between genes and complex traits. The detailed characterization of haplotype and LD profiles in candidate genes in different ethnic populations may shed light on population history and ethnic differences that may confound association studies, but can also facilitate the identification of variants that are most likely to be the primary etiological determinants of complex diseases (Johnson et al. 2001; Reich et al. 2001). Haplotype-based association studies have the advantage of not requiring the causal variant to be identified and tested directly and the potential of uncovering regions that harbor the trait variant (Judson and Stephens 2001). Such studies may provide more statistical power for detecting association with phenotype than association studies based on individual SNPs (Bader 2001). By identifying a subset of highly informative tagging SNPs (tSNPs) representing the most common variations segregating at the gene locus, haplotype-based association studies can also be performed more efficiently (Johnson et al. 2001).

This study postulates that both the *ABCC*4 and ABCC5 genes encoding widely distributed proteins with similarly unique transport profiles represent attractive candidate genes for linkage disequilibrium (LD) analysis and to test for signatures of local positive selection in preparation for future association studies. This study aims to derive population-specific haplotypes and generate tagging SNPs through comprehensive linkage disequilibrium analyses of *ABCC*4 and *ABCC*5 gene loci in 5 ethnically unique and geographically distinct populations. These tSNPs should simplify future disease/trait association studies examining the roles of *ABCC*4 and *ABCC*5. It will further assess if positive selection can be a major cause for the genetic

diversity observed in different populations at specific loci of *ABCC*4 or *ABCC*5. Loci exhibiting signatures of positive selection will be identified by a wide differentiation of allele frequency through two statistic measures, F_{st} and *P*excess (Akey et al. 2004; Bersaglieri et al. 2004), or the formation of a long-ranging haplotype within populations using a modified long-range haplotype test (LRH) (Sabeti et al. 2002; Tang et al. 2004).

II MATERIALS AND METHODS

Chapter 4 Genotype assay development

4.1 Multiplex PCR amplification

The multiplex minisequencing protocol is outlined in Figure 6. The five genomic segments containing the seven SNPs of MDR1 were amplified in a single multiplex PCR reaction. Twenty ng of genomic DNA was amplified in a T3 thermal cycler (Biometra) in a total volume of 10 μL containing 0.15 pmol/μL each of the 10 primers (Table 7A), 5 mM $MgCl_2$, 200 μM each of the four deoxynucleotide triphosphates (dNTPs), and 0.75 U of HotStarTaq polymerase in the PCR buffer that was supplied (Qiagen). The reaction mixture was subjected to initial denaturation at 94 °C for 15 min followed by 40 step-cycles of denaturation at 94 °C for 30 s, annealing at 56 °C for 30 s, and extension at 72 °C for 1 min. This was followed by a final extension at 72 °C for 5 min. The expected PCR fragments and their sizes are shown in Figure 7A.

4.2 Multiplex minisequencing

Unincorporated dNTPs and excess primers were inactivated and degraded in a single-step reaction by the addition of 5 U of exonuclease I and 0.5 U of shrimp alkaline phosphatase (SAP; United States Biochemical), respectively, to 1.3 μL of the PCR product in a final volume of 2 μL. The reaction mixture was incubated at 37 °C for 15 min, and the enzymes were subsequently inactivated at 80 °C for 15 min.

The treated PCR products were then subjected to a multiplex minisequencing reaction to interrogate the seven SNP loci simultaneously. SNP-specific probing primers (or minisequencing primers) were designed to anneal to template DNA next to each SNP site such that extension by DNA polymerase added a single dideoxyribonucleoside triphosphate (ddNTP) complementary to the nucleotide at the polymorphic site

(Figure 7B). Each of the four ddNTPs was labeled with a spectrally distinct fluorophore i.e. ddATP was labeled with dR6G, ddCTP with dTAMRA™, ddGTP with dR110, and ddTTP with dROX™. To facilitate the examination of the seven SNPs simultaneously, each SNP-specific primer was designed to be a different length by the addition of variable lengths of nonhomologous d(GACT) polynucleotide tails to the 5' end of the primer. This enabled differentiation of the SNP loci based on length of the different ddNTP-extended primers (Figure 7B). The addition of nonhomologous tails simplified the standardization of annealing temperatures for all primers regardless of their total primer lengths. Minisequencing primers longer than 40 bases were purified by HPLC to remove incomplete primer synthesis products. Table 7B details the sequences of the minisequencing primers and their concentrations in the final minisequencing reaction mixture.

The multiplex minisequencing reaction contained the treated multiplex PCR product (2 μL), various concentrations of minisequencing primers, and 1.3 μL of SNaPshot™ Multiplex Ready Reaction Mix (Applied Biosystems) in a total reaction volume of 5 μL. The reaction mixture was subjected to 25 single-base extension cycles of denaturation at 96 °C for 10 s, primer annealing at 53 °C for 5 s, and primer extension at 60 °C for 30 s. Thereafter, unincorporated fluorescent ddNTPs were inactivated enzymatically with 0.5 U of SAP at 37 °C for 1 h, followed by SAP deactivation at 80 °C for 15 min.

4.3 Capillary electrophoresis and data analysis

The multiplex minisequencing products (0.8 μL) were then mixed with 9 μL of HiDi™ formamide and 0.5 μL of GeneScan-120 LIZ size standard (Applied Biosystems), and resolved by automated capillary electrophoresis for 25 min on an

ABI PRISM 3100® Genetic Analyzer (Applied Biosystems). GeneScan-120 LIZ size standard was a 5th dye-labeled internal size standard. Analyses were performed with the GeneScan^{TM} 3.7 application software (Applied Biosystems). The relative position of each primer peak indicated the SNP locus, whereas the peak color(s) specified the genotype. The colors of the analyzed peaks were green for ddATP, black for ddCTP, blue for ddGTP, and red for ddTTP. Figure 7C provides a diagrammatic depiction of the various colored peaks denoting the alleles at the various SNP loci and their positions relative to one another.

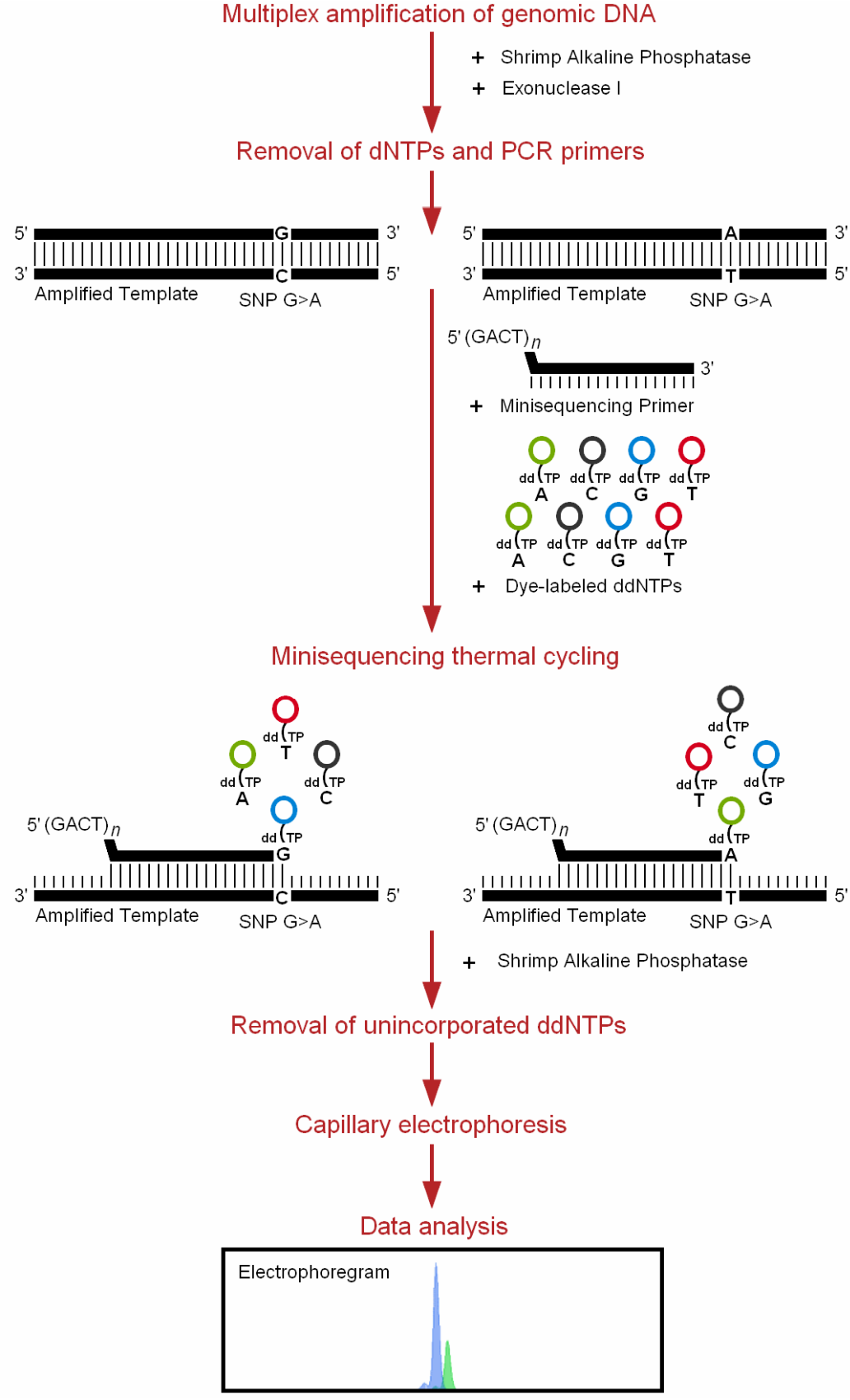


Figure 6. Schematic illustration of multiplex minisequencing protocol.

A. Conditions for Multiplex PCR of the 5 genomic conditions of the *MDR*1 gene

Amplified Region	*5'-Primer*	*3'-Primer*	*Conc (pmol/μL)*[a]	*Amplicon Length (bp)*
Exon 28	5'-TGGAAGAGGAATTAGGGAA-3'	5'-CCCACAAAAATGAGTAGGT-3'	0.15	761
5'UTR & Exon 1	5'-GGTGTTAGGAAGCAGAAAG-3'	5'-ACTATCCACGCCTCAAGA-3'	0.15	648
Exons 11 & 12	5'-TCTTTGTCACTTTATCCAGC-3'	5'-TCTCACCATCCCCTCTGT-3'	0.15	502
Exon 26	5'-CTCACAGTAACTTGGCAG-3'	5'-CTTACATTAGGCAGTGAC-3'	0.15	315
Exon 21	5'-TGCAGGCTATAGGTTCCAGG-3'	5'-TAGGGAGTAACAAAATAACAC-3'	0.15	284

B. Conditions for Multiplex Minisequencing of the 7 SNPs spanning 100 kb of the *MDR*1 gene

SNP	*Location*	*Region*	*Effect*	*Nucleotide Sequence*		*Primer orientation*	*Primer Sequence*[c]	*Length (bp)*	*Conc (pmol/μL)*	allele1 /migration length[d]	allele2/3 /migration length[d]
				Allele 1	*Allele 2/3*						
-41A>G	Intron -1	Promoter	N.A.	tccccaAtgattca	tccccaGtgattca	Reverse	5'AACGCGCATCAGCTGAATCA3'	20	0.04	A/26.6	G/24.3
-145C>G	Exon 1	Promoter	N.A.	aggaagCctgagc	aggaagGctgagc	Forward	5'*ACT(GACT)*$_7$CTCTCTTTGCCACAGGAAG3'	50	0.5	C/53.4	G/52.9
-129T>C	Exon 1	Promoter	N.A.	ttcgagTagcggc	ttcgagCagcggc	Reverse	5'*ACT(GACT)*$_3$*GAC*GAGCTTGGAAGAGCCGCT3'	36	0.15	T/40.9	C/39.5
1236T>C	Exon 12	NBD-1	Gly412Gly	gaagggTctgaac	gaagggCctgaac	Reverse	5'*CT(GACT)*$_8$*G*ACTCTGCATCTTCAGGTTCAG3'	56	1	T/59.8	C/58.7
2677G>T/A	Exon 21	TM-9, 10	Ala893Thr/Ser	gaaggtGctggga	gaaggtTctggga gaaggtActggga	Reverse	5'*T(GACT)*$_5$*GAC*TTAGTTTGACTCACCTTCCCAG3'	46	0.4	G/49.5	T/50.3 A/50.6
3435C>T	Exon 26	NBD2	Ile1145Ile	agagatCgtgagg	agagatTgtgagg	Reverse	5'*T(GACT)*$_{11}$*GA*CCTCCTTTGCTGCCCTCAC3'	66	0.07	C/68.0	T/69.0
4036A>G	Exon 28	3'UTR	N.A.	gaaatcAtagttt	gaaatcGtagttt	Forward	5'*CT(GACT)*$_2$TCATCAAGTGGAGAGAAATC3'	30	0.3	A/33.5	G/32.3

[a] Final concentration in the reaction mixture.

[b] UTR, untranslated region; N.A., not applicable.

[c] Italicized letters represent nonhomologous tail at 5′ end of SNP primer.

[d] Color of the allele represents SNP at particular loci, whereas migration length indicates the mean migration of that particular SNAPshot primer after primer extension and capillary electrophoresis.

Table 7. Multiplex PCR and minisequencing of the MDR1 gene.

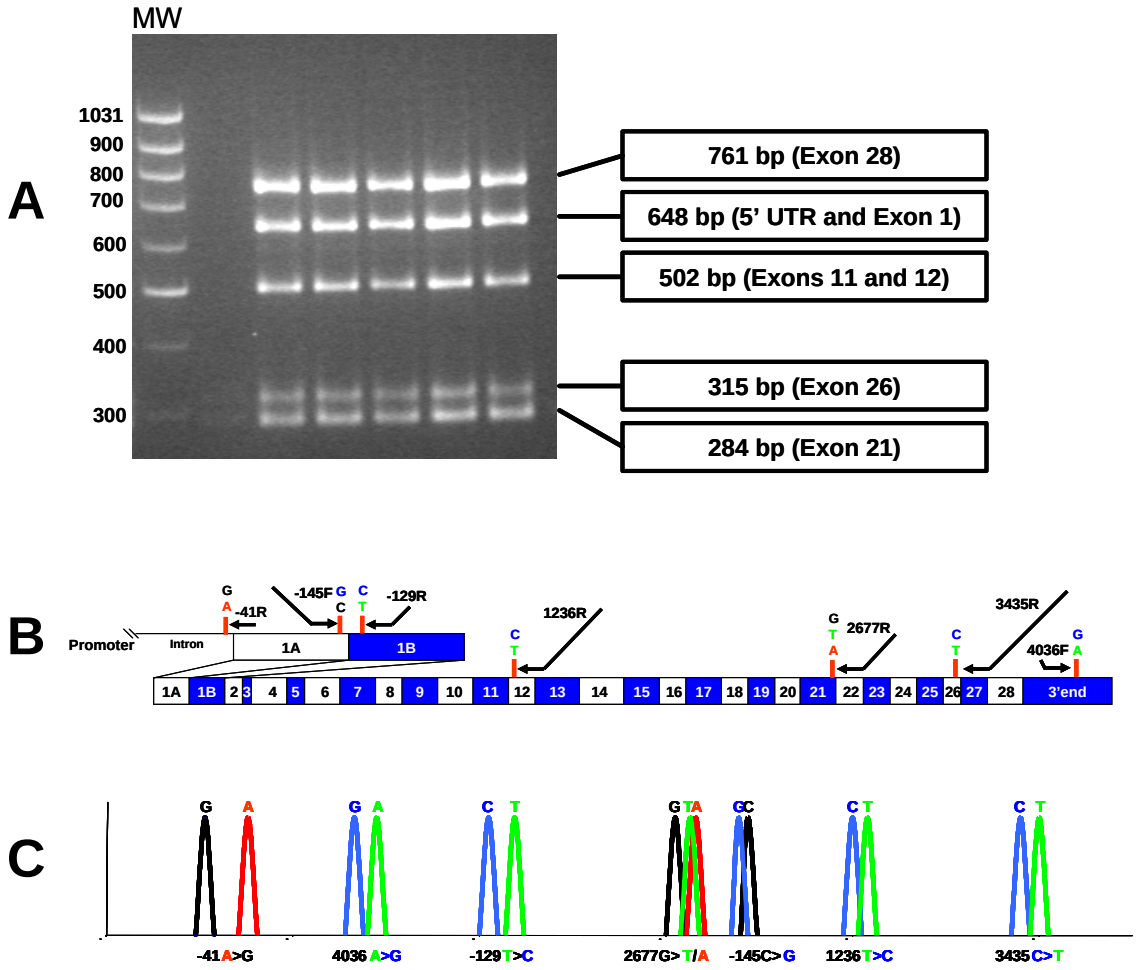


Figure 7. Multiplex PCR for the seven MDR1 SNPs.
Panel A agarose gel electrophoresis of the multiplex PCR products from five representative samples. A 2-µL aliquot of each reaction product was resolved on a 2% agarose gel. UTR, untranslated region.
Panel B, schematic diagram showing relative positions of the minisequencing primers and SNP sites within the MDR1 gene. Only exons are shown in the diagram except for the promoter region, where SNP -41AG resides within an intron. Minisequencing primers were designed to anneal next to each SNP site. Each minisequencing primer differed in length from the next by the inclusion of 5' nonhomologous tails of various lengths.
Panel C, schematic illustration of the expected allele peaks of the seven SNP loci and their relative migration patterns on the GeneScan electropherogram. The position of the ddNTP-extended primer peak specifies the SNP locus, whereas the peak color/fluorescence denotes the allele/nucleotide.

Chapter 5 Genetic structure characterization

5.1 DNA samples

Human genomic DNA was derived from 2 sources. Blood samples were collected from discarded umbilical cords of normal, healthy and unrelated neonates delivered in the National University Hospital, Singapore. Genomic DNA was isolated from 288 samples. These samples were previously archived genomic DNAs extracted from cord blood samples discarded after clinical newborn screening for glucose-6-phosphate dehydrogenase deficiency. The samples had been completely anonymized by removal of all identifiers immediately after G6PD testing and prior to DNA extraction. These samples belonged to each of three ethnic Asian populations representing various geographical locations including East Asia (Chinese) (96 samples), Southeast Asia (Malay) (96 samples) and South Asia (Indians) (96 samples). The Chinese and Indian populations were mainly descendents of migrants from South China and India respectively, while the Malays are indigenous inhabitants. Only sex and race of the sample are retained which are insufficient to determine identity. Ethnicity was categorized based on the mother's self-report and verified by the nurse. Individuals were considered to be of a given ethnic group if both parents were of the same ethnicity as the child. For example, the cord blood from a child of known mixed parentage would be assigned under ‘Others’ during childbirth and its cord blood would not be selected for DNA extraction. No database containing identifiable information exists for these samples, and there is thus no possibility of tracing their identity. These studies do not utilize prospectively collected venous blood samples. While efforts have been made to ensure careful handling during sample collection, there exists a small possibility that cord blood may be contaminated with maternal cells and thereby interfere with subsequent genotyping results. Nevertheless,

genotyping of microsatellites and the triallelic MDR1 SNP e26 2677G>T/A conducted on these same samples did not yield results of more than 2 alleles in any sample. This gives us confidence that any maternal DNA contamination is minimal among the samples. These studies fall within the guidelines as spelt out in the IRB guidelines (NUS_IRB guidelines (IRB-GUIDE-006 #4 and OHRP Guidelines 45 CFR 46.101) and the Human Tissue Research report of the Bioethics Advisory Committee (Part IV, Section 8, para 8.10) and exemption from IRB review was thus obtained from the National University of Singapore (NUS-IRB Reference Code #04-126E). All original identifiers of each sample were discarded with the exceptions of ethnicity and sex.

Genomic DNA samples were also obtained from the respective Human Variation Collections in the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository (The Coriell Institute for Medical Research, Camden, NJ). These consisted of 100 European Americans and 100 African Americans.

These populations were chosen with a view to directly compare against populations with similar ancestry used by the International Hapmap Project [http://www.hapmap.org/]. Hence the Chinese, European American and African American population data would be compared against data derived from the Han Chinese in Beijing (HCB), CEPH/UTAH (CEU) and Yoruba in Ibadan (YRI) respectively. The Malay and Indian populations were included to study whether results from one population could be transferable to other populations.

5.2 SNP selection

Selection of SNPs was initially based on three criteria: location at which a SNP could exert a functional change, high reported minimum allele frequency (>5%) for

adequate haplotype characterization and that consecutive SNPs were well-spaced apart throughout the *ABCC*4 locus for comprehensive coverage. Exonic SNPs are likely to cause functional changes, so all exonic SNPs that have a reported minor allele frequency of 5% were selected to maximize chances that these SNPs would be observed in the studied samples. SNPs in the 5' flanking region might alter promoter activity so these were studied as well. Intronic SNPs were selected purely to ensure an even SNP density throughout the gene locus and therefore a reported high minor allele frequency in intronic SNPs would be advantageous.

The *ABCC*4 gene (NT_009952) spans approximately 280 kb (Figure 8). In all, twenty-eight SNPs were selected from published reports (Saito et al. 2002a) and public databases (Figure 8). These SNPs lie within 1500 bp 5' from the transcription starting site to 1500 bp 3' downstream from the last exon of the gene. The distances between pairs of consecutive SNPs vary from 57 bp to 35.9 kb.

Twenty-one SNPs occurring in the exons, introns and flanking regions and spanning ~100 kb of the *ABCC*5 gene (NT_022959) were selected from previously published reports (Saito et al. 2002a; Dazert et al. 2003) or public databases (Figure 9). All reported SNPs residing in potentially functionally regions (i.e. exons and 5' flanking regions) were genotyped in at least one population. SNPs that were monomorphic in that population (e.g. exon 9 1200C>T; exon 25 3606C>A; exon 29 4148C>A; e30 5159C>T) were excluded from the final panel of SNPs that was genotyped in the 5 populations. The distances between 2 consecutive SNPs varied from 15 bp to 25 kb. In addition, 1163 bp upstream of the transcription start site was sequenced in more than 30 random Asian genomic samples (Figure 9) to identify additional novel SNPs in this very important putative promoter region. Three novel SNPs was identified in this region (Figure 9).

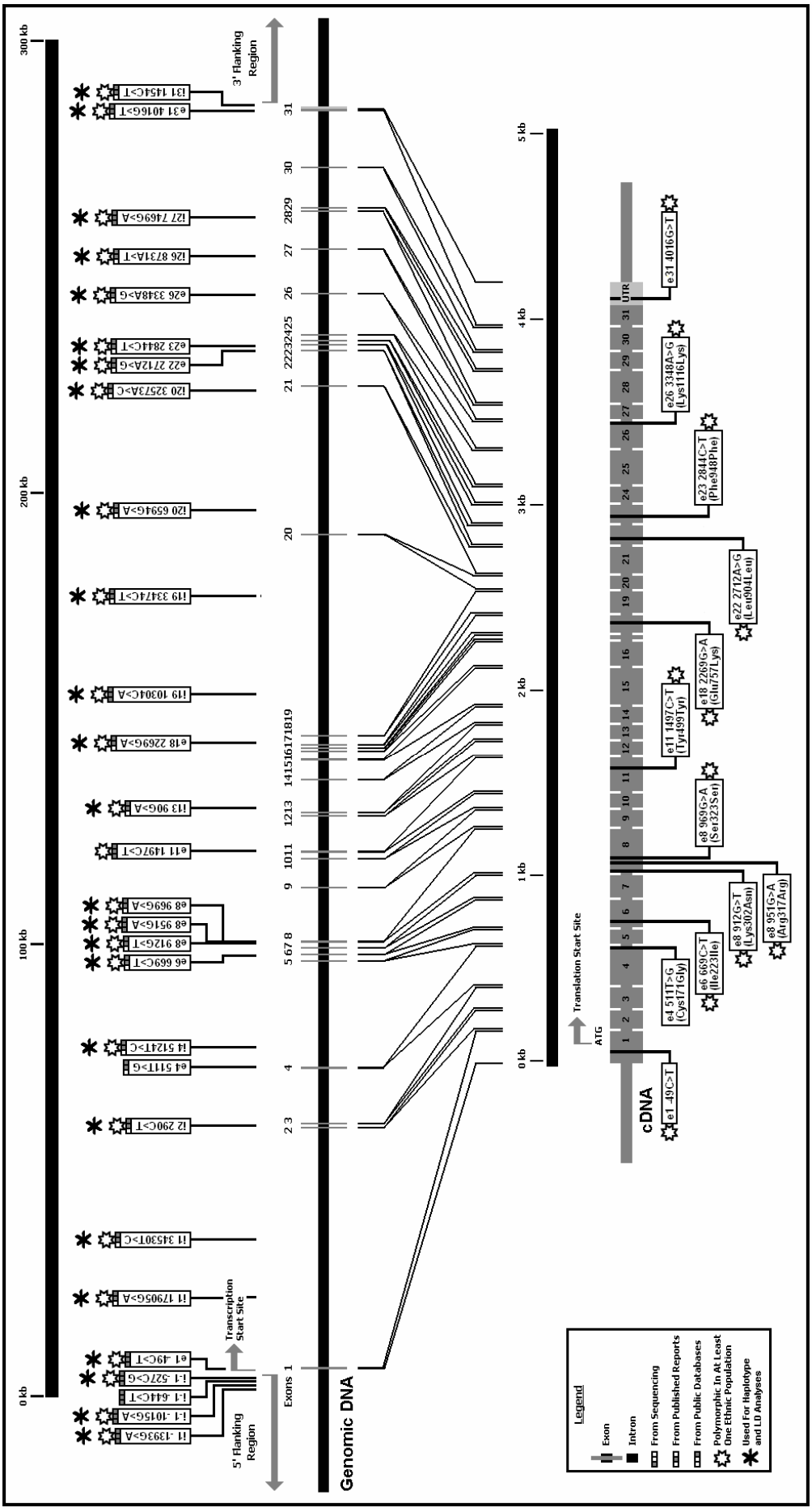


Figure 8. Distribution of SNPs across the *ABCC4* gene. The thick horizontal bar represents the entire genomic length of *ABCC4* with short vertical bars representing the exons. SNPs that are polymorphic in at least one population are marked with stars. SNPs that are selected for subsequent haplotype and LD analyses are marked with an asterisk. This map is drawn to scale according to the scale bar located above the genomic DNA or cDNA.

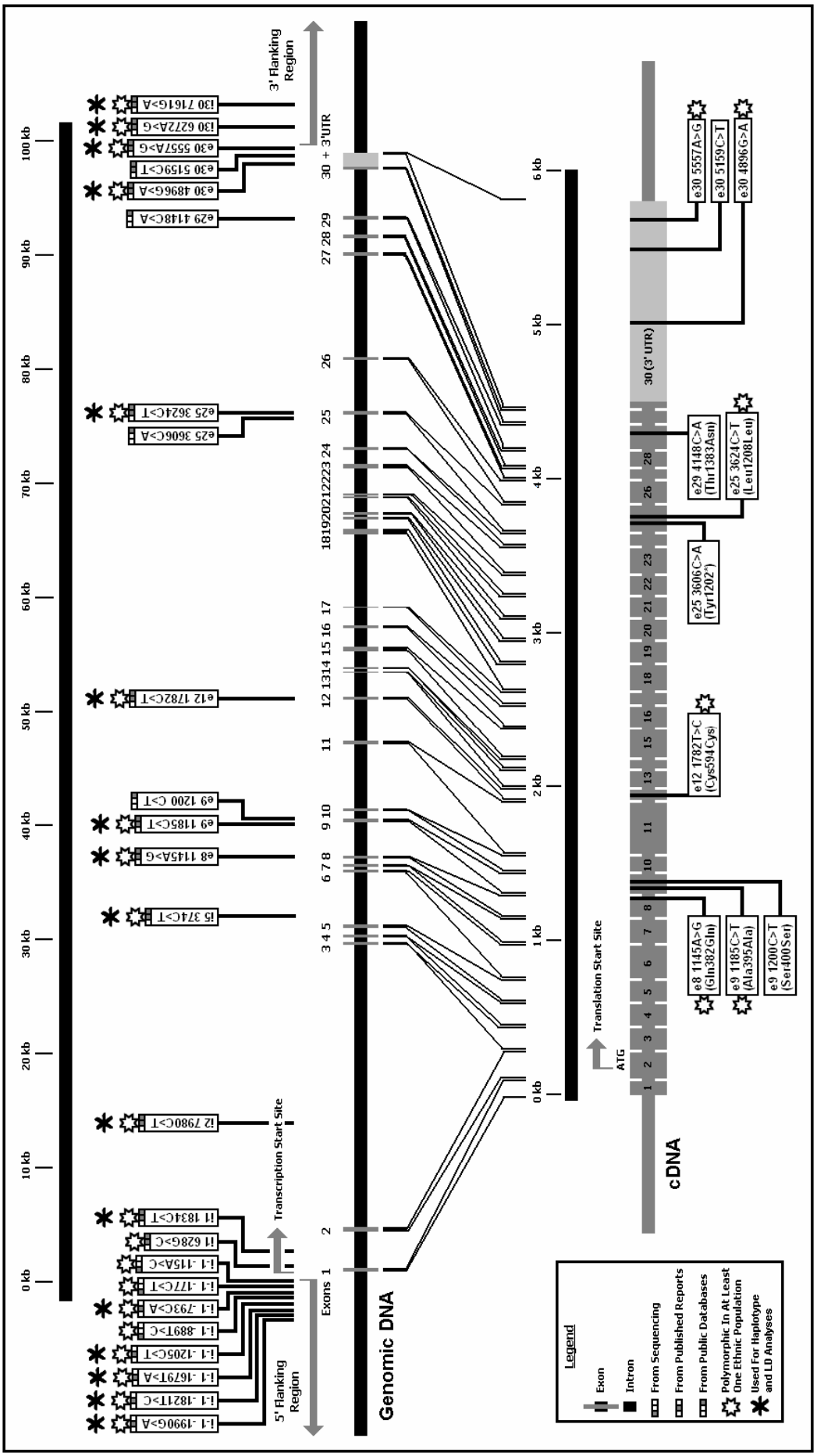


Figure 9. Distribution of SNPs across the *ABCC5* gene. The nomenclature for the SNPs is described in Methods. The thick horizontal bar represents the entire genomic length of *ABCC5* while short vertical bars represent the exons. SNPs that are polymorphic in at least one population are marked with stars. SNPs that are selected for subsequent haplotype and LD analyses are marked with an asterisk. This map is drawn to scale according to the scale bar located above the genomic DNA or cDNA.

5.3 Nomenclature of SNPs

The nomenclature for the SNPs used is as follows: For the 5' flanking SNPs - Region-number of nucleotides upstream the transcriptional starting site-Allele 1>Allele 2. Since these SNPs occur in the intronic region upstream exon 1, Region is denoted as intron -1 and allele numbers carry a negative sign. For the coding SNPs – Region-number of nucleotides downstream of transcriptional starting site using cDNA sequence-Allele 1>Allele 2. Region refers to the exon that the SNP resides. For 3' flanking SNPs – Region-sum of the length of the cDNA sequence plus number of nucleotides after the cDNA sequence-Allele 1>Allele 2. Since these SNPs occur in the intronic region after exon 30, Region is denoted as intron 30. For intronic SNPs – Region-number of nucleotides downstream the respective intron-exon junction-Allele 1>Allele 2. Region denotes the intron that the SNP resides (Gwee et al. 2005).

5.4 Genotyping strategy

Eight different minisequencing panels were set up, each genotyping 4-10 different *ABCC*4 and *ABCC*5 SNPs. Specific PCR primers and minisequencing probes were designed for the amplification of fragments containing these SNPs and the multiplex reactions respectively. Information on the PCR and minisequencing primers as well as the conditions of reaction are tabulated in the Tables 7 and 8. All primers ordered were normal oligonucleotides. Minisequencing primers that were greater than 45 bp in length were HPLC purified to reduce background and increase signal-to-noise ratio.

Genotyping was carried out using multiplex PCR amplification and multiplex minisequencing as previously described (Gwee et al. 2003). For each minisequencing panel, one to ten different genomic segments containing the SNPs of *ABCC*4 and *ABCC*5 were amplified in a single multiplex PCR reaction. PCR was performed

using either Qiagen Multiplex PCR Master Mix (containing HotStarTaq DNA Polymerase), or using standard HotStarTaq DNA polymerase in an otherwise identical reaction mixture. Ten ng of genomic DNA was amplified in a T3 thermal cycler (Biometra) in a total volume of 5 µL containing 0.10 – 0.25 pmol/µL each of the PCR primers (Table 7A), 1.25 - 5 mM $MgCl_2$, 100 – 200 µM each of the four deoxynucleotide triphosphates (dNTPs), and 0.25 - 0.75 U of HotStarTaq polymerase in the PCR buffer that was supplied (Qiagen). It was possible to substitute HotStarTaq polymerase directly with 2x Qiagen Multiplex PCR Master Mix (no additional buffer required). Multiplex assays could be tedious and time-consuming to establish. Some form of optimization was usually required, such as adjusting primer concentrations, and Mg^{2+} concentration. The Qiagen Multiplex PCR Kit combined the benefits of a highly stringent hot start with a unique PCR buffer specifically developed for multiplex reactions and made customization of multiplex PCR assays faster. For Panel C of Table 8 (*ABCC*4) and Panel of B of Table 9 (*ABCC*5), the high GC content (>65%) of the 5' and 3' flanking regions required the addition of betaine for successful amplification. First, the reaction mixture was subjected to initial denaturation at 95 °C for 15 min followed by 40 step-cycles of denaturation at 95 °C for 45 - 60s, annealing at 56 – 64 °C for 30 - 45s, and extension at 72 °C for 1 - 2 min. This was followed by a final extension at 72 °C for 10 min. To ensure that the multiplex amplification was successful, the amplicons were analyzed by polyacrylamide (10-15%) or agarose gel (1.5 - 2%) electrophoresis for at least 25% of the samples of each panel. The expected PCR fragments and their sizes are tabulated in Tables 7 and 8.

The direct addition of 2 U of exonuclease I and 0.2 U of shrimp alkaline phosphatase (SAP; United States Biochemical) to 0.5 µL of the PCR product in a final volume of

0.8 μL inactivated and degraded unincorporated dNTPs and excess primers respectively in a single-step reaction. The reaction mixture was incubated at 37 °C for 30 min, and the enzymes were subsequently inactivated at 80 °C for 15 min.

After treatment with ExoI and SAP, the PCR products were subjected to a multiplex minisequencing reaction. Each of the SNP-specific probing primers (or minisequencing primers) was designed to anneal to template DNA next to the respective SNP site so that DNA polymerase extended a single dideoxyribonucleoside triphosphate (ddNTP) complementary to the nucleotide at the polymorphic site. Incorporation occurred only at a single site and the addition of ddNTPs terminated further extension. Each of the four possible dye labeled terminators in the minisequencing reaction was labeled with a different fluorescent dye, which allowed the labeled primer extension products to be detected and analyzed. The addition of variable lengths of nonhomologous d(GACT) polynucleotide tails to the 5' end of each SNP-specific primer enabled differentiation of the SNP loci based on length of the different ddNTP-extended primers. The sequences of the minisequencing primers and their concentrations in the final minisequencing reaction mixture were tabulated in Tables 7 and 8 for *ABCC*4 and *ABCC*5 respectively. To 0.8 μL of treated multiplex PCR products, various concentrations of minisequencing primers and 0.28 μL of SNaPshot^{TM} Multiplex Ready Reaction Mix (Applied Biosystems) to a total reaction volume of 1.5 μL. The multiplex minisequencing reaction consisted of 25 single-base extension cycles of denaturation at 96 °C for 10 s, primer annealing at 53 °C for 5 s, and primer extension at 60 °C for 30 s. To inactivate unincorporated fluorescent ddNTPs enzymatically, 0.5 U of SAP was added to the tube. The entire solution was incubated at 37 °C for 30 min, followed by enzyme deactivation at 80 °C for 15 min. Before being resolved by automated capillary electrophoresis for 25 min on an ABI

PRISM 3100® Genetic Analyzer (Applied Biosystems), 0.8 μL of the multiplex minisequencing products were added to 9 μL of HiDi™ formamide and 0.15 μL of GeneScan-120 LIZ size standard (Applied Biosystems). Analyses of the electropherogram were performed with the GeneScan™ 3.7 application software (Applied Biosystems). For SNP loci in every panel of minisequencing, several samples from each ethnic population were randomly selected for dideoxy sequencing to ascertain the accuracy of genotyping and several samples were chosen to be regenotyped to ensure consistency of results obtained.

While the general steps of the genotyping strategies followed the original protocol used for haplotype analyses of MDR1 SNPs, several improvements had been made. These include the use of less DNA as starting material for amplification, reduction of volumes of reagents such as SNaPshot™ Multiplex Ready Reaction Mix (Applied Biosystems) as well as the reduction of time for amplification and minisequencing. This highlighted the flexibility and robustness of the genotyping strategy introduced.

Panel A

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
exon 4 511T>G	rs4148460	Exon 4	Cys171Gly	ttacgagtagccatg T gccatatgatttatc	ttacgagtagccatg G gccatatgatttatc	5'-AATGGTGTGGTCTGTATGTC-3'	5'-CAGGGTCTTGCTCTGTCAC-3'	0.200	406	E4/511M4F	Forward	5'-*C7GA* ATGAGGTTACGAGTAGCCATG-3'	25	0.200
exon 6 669C>T	rs899494	Exon 6	Ile223Ile	accactgcaggcgat C gcagtgactgcccta	accactgcaggcgat T gcagtgactgcccta	5'-CAGGTGTAAACCCACTGTG-3'	5'-GAAGACTGTGGATGTTACC-3'	0.400	594	E6/669M4R	Reverse	5'-CAGAGTAGGGCAGTCACTGC-3'	20	0.050
exon 8 912G>T	rs2274407	Exon 8	Lys302Asn	tgtactctctttcag G aaggagatttccaag	tgtactctctttcag T aaggagatttccaag	5'-GCGTGAGCCACTTTATCTG-3'	5'-GATAGGGAAGTACACACAAC-3'	0.200	738	E8/912M4F	Forward	5'-*GACTGAC* TAAAAACCTGTACTCTCTTTCAG-3'	30	0.400
exon 8 969G>A	rs2274405	Exon 8	Ser323Ser	gatgaatttggcttc G tttttcagtgcaagc	gatgaatttggcttc A tttttcagtgcaagc	5'-GCGTGAGCCACTTTATCTG-3'	5'-GATAGGGAAGTACACACAAC-3'	0.200	355	E8/969M4F	Forward	5'-*T(GACT)$_8$GA* CAGAGGGATGAATTTGGCTTC-3'	44	0.800
exon 11 1497C>T	rs1557070	Exon 11	Tyr499Tyr	atttgggaagaaata C gaaaaggaacgatat	atttgggaagaaata T gaaaaggaacgatat	5'-GTGGCTTATTCTTGTCTCG-3'	5'-CTAGTATTACTGGACATTCG-3'	0.200	466	E11/1497M4R	Reverse	5'-*ACTGACTGA* GACTTTTTCATATCGTTCCTTTTC-3'	33	0.100
exon 18 2269G>A	rs3765534	Exon 18	Glu757Lys	ggaggaaatgtaacc G agaagctagatctta	ggaggaaatgtaacc A agaagctagatctta	5'-TTTGTCTTTGTTGTTTCTCCC-3'	5'-CTCAGCCTCTCATACAATAC-3'	0.200	301	E18/2269M4F	Forward	5'-*T(GACT)$_4$G* AAATGGAGGAGGAAATGTAACC-3'	40	0.100
exon 22 2712A>G	rs1678339	Exon 22	Leu904Leu	agtgttttcccactt A tcatcttctctccag	agtgttttcccactt G tcatcttctctccag	5'-GCACCCAGCCTCACTGAG-3'	5'-GATCCTGTCCTTTGAAACC-3'	0.200	202	E22/2712M4R	Reverse	5'-*C7(GACT)$_5$GA* AGCCCCTGGAGAGAAGATGA-3'	48	0.400
exon 23 2844C>T	rs1189466	Exon 23	Phe948Phe	aacgtcccgctggtt C gccgtccgtctggat	aacgtcccgctggtt T gccgtccgtctggat	5'-3'TCATGCTTATTACCGTTCTC-3'	5'-CATAGTTATTCACATCTGTAC-3'	0.200	164	E23/2844M4R	Reverse	5'-*T(GACT)$_9$* TGGCATCCAGACGGACGGC-3'	56	0.200
exon 26 3348A>G	rs1751034	Exon 26	Lys1116Lys	cgatttaaggaagaa A atgtcaatcatacct	cgatttaaggaagaa G atgtcaatcatacct	5'-CTGCCGCCTGGATCTCTC-3'	5'-AAAAGCAATTAAACATAGTAGC-3'	0.400	242	E26/3348M4F1	Forward	5'-*(GACT)$_{11}$G* GGACTTCACGATTTAAGGAAGAA	64	0.100
exon 31 4016G>T	rs3742106	Exon 31	3' UTR	gttccgaaggcattt G ccactagtttttgga	gttccgaaggcattt T ccactagtttttgga	5'-CACATGGTTACAAACACTTCC-3'	5'-ACCTGATAGACGGCATTTAAC-3'	0.400	323	E31/4016M4F	Forward	5'-*C7(GACT)$_{12}$G* AAGTCCGTTCCGAAGGCATTT-3'	72	0.800

Panel B

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
intron 1 34530T>C	rs871175	Intron 1		agggataatgcccca T ttactaatgggggaa	agggataatgcccca C ttactaatgggggaa	5'-GTAGGGACACAACCAAACC-3'	5'-CACCAACAAGCGACAGGC-3'	0.200	164	I1/34530M4F	Forward	5'-*C7(GACT)$_7$GA* CACCAAAGGGATAATGCCCCA-3'	53	0.200
intron 4 5124T>C	rs2389234	Intron 4		tgccaggtaaagcta T agcagtacctgagcc	tgccaggtaaagcta C agcagtacctgagcc	5'-CAGTGGGTTGGCTTCTGAG-3'	5'-CTAAAGAAAGTGAAGACGAGC-3'	0.200	192	I4/5124M4F	Forward	5'-*C7(GACT)$_2$GAC* TCCTGAATGCCAGGTAAAGCTA-3'	35	0.600
intron 13 90G>A	rs1751005	Intron 13		ctttcagctcatcct G ctctggtgtccgctc	ctttcagctcatcct A ctctggtgtccgctc	5'-GCTGACATCTATCTCCTGG-3'	5'-CCTTGACCTTGTATTTCTGC-3'	0.200	249	I13/90M4R	Reverse	5'-GGGGAGCGGACACCAGAG-3'	18	0.320
intron 19 33474C>T	rs1189429	Intron 19		tagagaatggcacca C aagggcatgagatca	tagagaatggcacca T aagggcatgagatca	5'-AGTGACAGTTATTGAGGTTTC-3'	5'-ATCTGCCTCTTTCCACTCC-3'	0.200	290	I19/33474M4R	Reverse	5'-*T(GACT)$_4$GA* CAGTGGTGATCTCATGCCCTT-3'	40	0.056
intron 20 6594G>A	rs2766481	Intron 20		cctttaaggaccagc G aaagtctctctttcg	cctttaaggaccagc A aaagtctctctttcg	5'-TACAAAGGACATACAAAGAGC-3'	5'-TAAATGAGATGGGGCAAGTG-3'	0.300	341	I20/6594M4R	Reverse	5'-*(GACT)$_4$GA* CACTAATGCGAAAAGAGAGACTTT-3'	30	0.640
intron 20 32573A>C	rs1189437	Intron 20		taacaacaactcatg A aagtatttttgaaaa	taatcctaacaacaactc atg C aagtatttttgaaaa	5'-ACACCATCTCTACTAAAAATAC-3'	5'-CAGAGACCACACCAACCAC-3'	0.300	457	In20/32573M4F	Forward	5'-*ACTGACT* TGTAATCCTAACAACAACTCATG-3'	44	0.320
intron 26 8731A>T	rs1211485	Intron 26		ttttttttttttttgaga A ggagttttgctcttg	ttttttttttttttgaga A ggagttttgctcttgttg	5'-AGCAGATGAATAGGAACAGC-3'	5'-ACCAACATAG˜GAAACCCTG-3'	0.300	503	I26/8731M4R	Reverse	5'-*(GACT)$_6$GAC* GGGCAACAAGAGCAAAACTCC-3'	48	0.140

Table 8. Primers, PCR and Minisequencing Conditions for 28 SNPs at the *ABCC*4 gene locus.

Panel C

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Mini-sequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
intron -1 -1393G>A	rs868853	5' Flanking		ttcctcatctataaa G taaggataaccttga	ttcctcatctataaa A taaggataaccttga	5'-CCTACAGCCATTCAACCAG-3'	5'-GTGACCTGTTTCGGGCGG-3'	0.900	1780	I-1/-1393M4F	Forward	*5'-CT(GACT)$_3$GAC* GCTTCACCTTCCTCATCTATAAA-3'	40	0.920
intron -1 -1015G>A	rs2993579	5' Flanking		ctagctacttaaact G agttgggaggattgc	ctagctacttaaact A agttgggaggattgc	*As Above*	*As Above*	*As Above*	*As Above*	I-1/-1015M4R	Reverse	*5'-(GACT)$_5$GAC* GCTCAAGCAATCCTCCCAACT-3'	44	0.800
intron -1 -644C>T	rs3814270	5' Flanking		tctgggtcatactct C gagttacccggcttt	tctgggtcatactct T gagttacccggcttt	*As Above*	*As Above*	*As Above*	*As Above*	I-1/-644M4F	Forward	*5'-T(GACT)$_6$G* GAATTCATCTGGGTCATACTCT-3'	48	0.800
intron -1 -527C>G	rs869951	5' Flanking		gcctcccatggcacc C tcgtttggtcctgag	gcctcccatggcacc G tcgtttggtcctgag	*As Above*	*As Above*	*As Above*	*As Above*	I-1/-527M4R	Reverse	*5'-(GACT)$_2$GA* CCCTTCTCAGGACCAAACGA-3'	30	0.400
exon 1 -49C>T	rs3751333	Exon 1		ggagcccgcggccac C gccgcctgatcagcg	ggagcccgcggccac T gccgcctgatcagcg	*As Above*	*As Above*	*As Above*	*As Above*	E1/-49M4R	Reverse	*5-T(GACT)$_4$* TCGCGCTGATCAGGCGGC-3'	35	0.400

Panel D

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
intron 1 17905G>A	rs9524885	Intron 1		cttcaaggtggctct G ggattagcagcatca	cttcaaggtggctct A ggattagcagcatca	5'-AGGAATGGAGGGAATGAGTT-3'	5'-CTTGTAGAACGTGATCAAAATG-3'	0.800	637	I1/17905M4R	Reverse	5'-GCTGCTGATGCTGCTAATCC-3'	20	0.200
intron 2 290C>T	rs4148436	Intron 2		gttgtctggcttatc C gtgctactgatggg	gttgtctggcttatc T gtgctactgatgggg	5'-GGTGTCTGTGTGTGGGGAGT-3'	5'-TAAGAGTGAACCCTGCCACA-3'	0.150	482	I2/290M4F	Forward	*5'-CTGACTGAC* GAGGAGGTTGTCTGGCTTATC-3'	30	0.600
exon 8 951G>A	rs2274406	Exon 8	Arg317Arg	aagttcctgcctcag G gggatgaatttggct	aagttcctgcctcag A gggatgaatttggct	5'-TTTATCTGGTTGACATCACTGC-3'	5'-GCAGTTACCACCTGTACATCA-3'	0.200	558	E8/951M4R	Reverse	*5'-CT(GACT)$_4$GAC* AAAAATGAAGCCAAATTCATCCC-3'	48	0.400
intron 19 10304C>A	rs1189429	Intron 19		ataatctctaagctt C aagtggactcacaac	ataatctctaagctt A aagtggactcacaac	5'-ACAACACACATGAGCATAGC-3'	5'-AGGTACTTGTAAAGGGCAAC-3'	0.150	351	I19/10304M4R	Reverse	*5'-CT(GACT)$_2$GA* TAGAAATTGTTGTGAGTCCACTT-3'	35	0.200
intron 27 7469G>A	rs1151471	Intron 27		cccaatgtagtatca G atcatcttatagcta	cccaatgtagtatca A atcatcttatagcta	5'-TGACTCTTGGTTCCTCTATAGC-3'	5'-AGGACACAATATAACATCTTGC-3'	0.150	202	I27/7469M4F	Forward	*5'-ACT(GACT)$_3$GAC* ACCACCACCCAATGTAGTATCA-3'	40	0.320
intron 31 1454C>T	rs1059762	3' Flanking		tgaggtttgttaaaa C gatagctatattcaa	tgaggtttgttaaaa T gatagctatattcaa	5'-CCTCTCAGAATAAGGTGTCAC-3'	5'-CATTAAAACAGAAACAGGACG-3'	0.400	437	I31/1454M4F	Forward	*5'-CT(GACT)$_4$G* GTGATCATAATGAGGTTTGTTAAAA-3'	44	0.480

Panel A

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
e9 1185C>T	rs1132776	Exon 9	Ala395Ala	tggaaaaagc C gggtacttcc	tggaaaaagc T gggtacttcc	5'-TTGCTTTGAAATTGGCTTGC-3'	5'-CCATCCCTGAGGGTTCCC-3'	0.220	204	e9/1185M5F	Forward	5'-CGTCGGATATTGGAAAAAGC-3'	20	0.100
e12 1782c>T	rs939336	Exon 12	Cys594Cys	ttggaatctg C ggcagtgtgg	ttggaatctg T ggcagtgtgg	5'-TCTGTGTGCCTTGTCCAG-3'	5'-AAACCAGAGAAAGCAGCAG-3'	0.220	134	e12/1782M5F	Forward	5'-*(GACT)*$_4$ GGTAAACTGGTTGGAATCTG-3'	36	0.200
e25 3624C>T	rs3749442	Exon 25	Leu1208Leu	agatgaggta C cgagaaaacc	agatgaggta T cgagaaaacc	5'-TGGTCAGTCTGAACTGGC-3'	5'-GCTGAGACACTAATTGCTC-3'	0.160	340	e25/3624M5R	Reverse	5'-*ACT(GACT)*$_6$ AGAACGCAGAGATGAGGTA-3'	30	0.200
e30 4896G>A	rs3749445	Exon 30		cgcctccccac G gccgctcca	cgcctccccac A gccgctcca	5'-CGTGTGGCAATAGTGGGC-3'	5'-GCTAAGGCGACAGAATGTC-3'	0.200	764	e30/4896M5F	Forward	5'-*ACT(GACT)*$_5$ CTCTGCCGCCTCCCCAC-3'	44	0.200
e30 5557A>G	rs562	Exon 30		gctgtcatca A ttgaatggtc	gctgtcatca G ttgaatggtc	*As Above*	*As Above*	*As Above*	*As Above*	e30/5557M5R	Reverse	5'-*(GACT)*$_9$*GAC* CATGCAACGCTGACCATTCAA-3'	60	0.600

Panel B

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
i-1 -889T>C		5' Flanking		aggcggacgc T gcagtgagcc	aggcggacgc C gcagtgagcc	5'-GGAGAACTGCTTGAACCCG-3'	5'-CCAAGGCTGAAGGGAACTG-3'	0.200	851	i-1/-889M5R	Reverse	5'-*CT(GACT)*$_8$*GC* GATCTCGGCTCACTGC-3'	52	0.200
i-1 -177C>T		5' Flanking		agaggggcag C ccgcaccgcc	agaggggcag T ccgcaccgcc	*As Above*	*As Above*	*As Above*	*As Above*	i-1/-177M5F	Forward	5'-*GA* CGGTGGGAAGAGGGGCAG-3'	20	0.560
i-1 -115A>C		5' Flanking		ccgcgtccgc A gccctgggca	ccgcgtccgc C gccctgggca	*As Above*	*As Above*	*As Above*	*As Above*	i-1/-115M5F	Forward	5'-*ACT(GACT)*$_2$*GA* TCCTTCGCCGCGTCCGC-3'	30	0.800
i30 6272A>G	rs1000002	3' Flanking		tgctgtccaa A tgagaaccat	tgctgtccaa G tgagaaccat	5'-TGCTAGATCAGATTGCCTC-3'	5'-TCAAATGGTAGGTGGCTGG-3'	0.400	1301	i30/6272M5R	Reverse	5'-*ACT(GACT)*$_4$ TACTAGCTCTGATGGTTCTCA-3'	40	0.400
i30 7161G>A	rs1533682	3' Flanking		gtcaacagaa G ggttttgcta	gtcaacagaa A ggttttgcta	*As Above*	*As Above*	*As Above*	*As Above*	i30/7161M5R	Reverse	5'-*(GACT)*$_9$ GAGGCTGCTCTAGCAAAACC-3'	56	1.600

Panel C

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
i1 628G>C	rs4148556	Intron 1		cagagccgcg G gtggctttgt	cagagccgcg C gtggctttgt	5'-CTCTCCACGGTCATCCAG-3'	5'-TTCTAACCTTCCCCTTCAAG-3'	1.200	1368	i1/628M5R	Reverse	5'-*CTGACTGAC* GTGTGATAAACACAAAGCCAC-3'	30	4.000
i1 1834C>T	rs4148557	Intron 1		tgacctcctc C gtttcaaact	tgacctcctc T gtttcaaact	*As Above*	*As Above*	*As Above*	*As Above*	i1/1834M5R	Reverse	5'-*ACT(GACT)*$_3$*GA* GCGGTGAGCAGTTTGAAAC-3'	36	4.000
i2 7980C>T	rs2292997	Intron 2		tccctgtagg C atccataaag	tccctgtagg T atccataaag	5'-CAGGAGGAAGGAGATTAGC-3'	5'-ACCAAGCATGAGAGCCACC-3'	1.200	281	i2/7980M5F	Forward	5'-TTTTCTTCTCTCCCTGTAGG-3'	20	0.200
i5 374C>T	rs3749438	Intron 5		gagccagcac C gggaacatac	gagccagcac T gggaacatac	5'-GGGTGAAATGGAACTGACTC-3'	5'-GGTGCCCAGGAAACAGAG-3'	1.200	390	i5/374M5R	Reverse	5'-*(GACT)*$_6$*G* AGGCACTTGGTATGTTCCC-3'	44	0.150

Table 9. Primers, PCR and minisequencing conditions for 20 SNPs at the *ABCC5* gene locus.

Panel D

SNP	NCBI SNP ID	Location	Amino Acid Change	Nucleotide Sequence		PCR Conditions				Minisequencing Conditions				
				Allele 1	Allele 2	5' Primer	3' Primer	Concentration (pmol/ul)	Amplicon Length (bp)	Minisequencing Primers	Primer Orientation	Sequence	Total Primer Length (bp)	Concentration (pmol/ul)
i-1 -1990G>A		5' Flanking		agctgggtgc G gtggccggtg	agctgggtgc A gtggccggtg	5'-CAACATATATGAAGTATTTCAGCGG-3'	5'-GAAATATCTTTATGAACTTGGGAG-3'	0.900	770	i-1/-1990M5R	Reverse	5'-GACTACAGACACCGGCCAC-3'	19	0.060
i-1 -1821T>C		5' Flanking		atttccctga T tatgaatgag	atttccctga C tatgaatgag	*As Above*	*As Above*	*As Above*	*As Above*	i-1/-1821M5R	Reverse	*5'-(GACT)*$_3$ AAAGACACCCAATCTCATTCATA-3'	35	0.600
i-1 -1679T>A		5' Flanking		attttgaata T ttattattca	attttgaata A ttattattca	*As Above*	*As Above*	*As Above*	*As Above*	i-1/-1679M5F	Forward	*5'-(GACT)*$_3$*G* TACTGAACTTAGAAATTATTTTGAATA-	40	1.200
i-1 -1205C>T		5' Flanking		aactagctag C tgttataaag	aactagctag T tgttataaag	5'-CAAGAATATGCTGCTTTACG-3'	5'-TAACCCGTTGAGAGTCGTCA-3'	0.900	859	i-1/-1205M5R	Reverse	*5'-ACTGACT* GTGACTTGTCCATCTTTATAA-3'CA	30	0.640
i-1 -793C>A		5' Flanking		actcataggg C tgtaatggag	actcataggg A tgtaatggag	*As Above*	*As Above*	*As Above*	*As Above*	i-1/-793M5R	Reverse	*5'-(GACT)*$_5$ TTGTTTCCTTTAATCTCCATTACA-3'	44	1.000
e8 1145A>G		5' Flanking	Gln382Gln	agagtgttca A agtgagttta	agagtgttca G agtgagttta	5'-TTGAATGAGCTTTAATTTGGCG-3'	5'-GCAAAGTGTGTGAATTATAAGACC-3'	0.200	269	e8/1145M5F	Forward	*5'-GAC* AAGCATTTTCTCAGAGTGTTCA-3'	25	0.100

Table 9 (continued).

5.5 Data analysis

The allele frequencies for each SNP were calculated and determined if these frequencies vary significantly among 5 populations using Fisher's exact test. Any deviation of the SNP allele frequency from Hardy-Weinberg equilibrium was also tested using Bonferroni correction for multiple locus testing.

All SNPs polymorphic in all populations were used for the inference of haplotypes from the genotypic data, removing SNPs that were monomorphic in any population. Only samples in which all loci can be genotyped were included in the haplotype frequency estimation. To reconstruct possible haplotypes and estimate haplotype frequencies from genotype data, the expectation-maximization (EM) algorithm (Slatkin and Excoffier 1996) within SNPHAP version 1.3 program [http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt] was utilized. This program has the advantage of managing potentially large number of haplotypes generated from a large number of SNP loci. Fisher's exact test was used to detect any differences in haplotype frequencies within populations.

To summarize pair-wise LD, 2 common measures of LD, Lewontin's coefficient, $|D'|$ (Lewontin and Krakauer 1973; Ardlie et al. 2002), and Pearson's correlation, r^2 (Pritchard and Przeworski 2001; Ardlie et al. 2002) were used. Fisher's exact test was utilized to determine whether pairs of SNP markers were in significant LD. From the negative correlation of LD and genetic distance, a LD decay trend line was obtained with the equation: $D = Do(1-\alpha)t$ (Schulze et al. 2002). Half-LD ($LD_{0.5}$) is defined as the distance at which $|D'|$ is 0.5.

5.6 Identification of tagging SNPs

Tagging SNPs (tSNPs) at the both the *ABCC*4 and *ABCC*5 gene locus were identified using the "TagIT" program [http://popgen.biol.ucl.ac.uk/software.html] based on weighted-average haplotype r^2 (Weale et al. 2003). The smallest 'best-performing' sets of tagging SNPs in every population as well as for all 5 populations were chosen.

5.7 Scanning for evidence of positive selection

Two measures, F_{st} and *P*excess based on the quantitation of the variation in SNP allele frequencies between populations were used to identify signatures of natural selection. To examine variations in allele frequencies, distribution of F_{st} values was constructed for all genotyped SNPs for 5 populations. The fixation index, F_{st}, was calculated with Nei's correction for sample size (Akey et al. 2002). In cases where estimates of F_{st} values are negative, they are assumed to be zero (Akey et al. 2002). Based on the comparison of simulated and empirical genome-wide distribution of F_{st} performed by Akey et al (Akey et al. 2002), a threshold of 0.45 or greater was applied in this study to be indicative of selective forces acting on that allele.

Besides F_{st}, a novel test based on *P*excess was applied. This test calculates the probability P which represents the extent at which a haplotype is over-represented in the affected sample, and equates to **P**excess = (**P***test* - **P***reference*)/(1 - **P***reference*), where **P***test* and **P***reference* denote allele frequency in chromosomes of population under study and ancestral population, respectively. Taking the African American allele frequency as representative of the ancestral allele frequency (**P***reference*), the 21% European admixture was not corrected (Parra et al. 1998; Bersaglieri et al. 2004). As a final strategy to look for signatures of recent positive selection at both the nucleotide analogue transporters, this study also adopted the long range haplotype

(LRH) test described first by Sabeti et al., 2002 and modified it for individual SNP markers. The Haplotype Branching Diagrams (HBDs) were illustrated using an algorithm that was programmed using the VBA language in Microsoft Excel. The algorithm was capable of analyzing both biallelic and triallelic SNPs. The Haplotype Branching Diagram (HBD) allowed the profile of LD decay to be represented pictorially across the entire *ABCC*4 or *ABCC5* gene locus and helped the visualization of the presence of a strong and extensive haplotype, signified by a strong thick branch. The root of each diagram begins from a core SNP locus from which branches appear. The thickness of each branch corresponds to the proportion of chromosomes with the haplotype.

To quantitatively assess the LD decay of each allele (for every test SNP site) over a stretch of genomic distance, a statistical measure called the extended haplotype homozygosity (EHH) was used. EHH at a distance x from the test variant (either a single SNP allele or a core haplotype) is defined as the probability that two randomly chosen chromosomes containing the tested variant are homozygous at all loci for the entire length of x. Relative EHH (rEHH) is then expressed as the ratio of EHH of the tested allele against the EHH of the other alleles at the same locus (Sabeti et al. 2002). A high relative rEHH given a certain allele frequency is usually indicative of recent positive selection. The extended haplotype homozygosity (EHH) of each tested SNP allele was first calculated to capture those with unusually high rEHH. The rEHH values of these alleles were then tested against simulated data.

5.8 Coalescent simulations

An allele with long flanking haplotype may be under evolutionary processes other than recent positive selection. To test the hypothesis of positive selection, coalescent

simulations were performed (Hudson 2002). Representing different evolutionary processes, four different population models including those of constant size, expansion, extreme bottleneck and highly structured population were simulated. 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 reduction in population size 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 with a compound term signifying a constant migration rate $4N_e\mu=0.1$ was assumed for the structured population, where μ is the mutation rate.

Within each model, simulations were performed with three different recombination rates, 0.65 cM Mb^{-1}, 1.3 cM Mb^{-1} and 2.6 cM Mb^{-1}. These values reflect the large variations of recombination rates observed in the human genome (Nachman 2002). Three mutation rates, $0.5x10^{-8}$ $site^{-1}$ $generation^{-1}$, $1.0x10^{-8}$ $site^{-1}$ $generation^{-1}$ and $2.0x10^{-8}$ $site^{-1}$ $generation^{-1}$, were first tested for the constant-size model with recombination rate of 1.3 cM Mb^{-1}. The distribution of mutation sites were expected to be independent of genealogy, since these mutation sites were assigned after simulation of genealogy. The resulting data distribution was found to be consistent despite different mutation rates. Thus, a constant mutation rate of $1.0x10^{-8}$ $site^{-1}$ $generation^{-1}$ was utilized for all models. Only recombination rates would be variable for each of the 4 models.

A chromosomal sequence of 300 kb was simulated. Datasets matching the observed data to within ±12.5% of the allele frequency and EHH, and within ±3 kb in distance for all downstream loci were first selected. The anchor locus was chosen to be within

±3 kb of the tested locus for the selected datasets. Each simulation was iterated at least 10,000 times to achieve 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 of alleles of the test SNP locus. 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 ±0.25 allele frequency window of that SNP. SNPs of high rEHH were selected as candidates for ranking against simulated data points containing SNPs of the matching allele frequency. SNP markers with empirical P value < 0.05 were deemed to show evidence of recent positive selection.

5.9 Prediction of transcription factor binding sites

To assess whether the identified polymorphisms in 5' flanking region (*ABCC*4 SNP -1 -1015G>A) affected binding to putative transcription factors, a genomic sequence of 1500 bp 5' from the transcription starting site containing each of the alternative alleles of SNPs within this sequence was submitted to three web-based algorithms. These three webtools were MatInspector [http://www.genomatix.de/products/MatInspector/index.html] (Quandt et al. 1995), TESS (Transcription Element Search System) [http://www.cbil.upenn.edu/tess] (Schug and Overton 1997) and Alibaba2.1 [http://www.alibaba2.com/] (Grabe 2002). The MatInspector algorithm was set with the default optimized matrix similarity threshold and produced the most number of results. The TESS algorithm was set at these default parameters: maximum allowable string mismatch % (t_{mm}) = 10;

minimum log-likelihood ratio score (t_{s-a}): = 12; minimum string length (t_w) = 6; secondary lg-likelihood deficit = 3; count significance threshold = $1.0e^{-2}$; minimum core similarity = 0.75; and minimum matrix similarity = 0.85. These parameters were set for the Alibaba2.1 algorithm: pair-wise similarity of sequence to known binding sites = 50; matrix width in bp = 10; minimum number of sites in matrix = 4; average matrix conservation = 75%; Similarity of sequence to matrix = 1%; and factor class level = 4.

III RESULTS

Chapter 6 Genotyping assay development

The aim of this study was to use the MDR1 gene as a candidate for the development of a genotyping assay to be used for haplotype analysis and this study will serve as a foundation for the genetic characterization of the nucleotide analogue transporters *ABCC*4 and *ABCC*5, the subjects of this thesis.

A multiplex PCR strategy was employed to simultaneously amplify 5 different genomic regions in the *MDR*1 gene for subsequent genotyping of the 7 SNPs in the *MDR*1 gene. Under the PCR conditions described, 5 distinct bands of different sizes representing the 5 genomic regions were observed (Figure 7).

The multiplexed PCR products were then purified to inactivate and remove unincorporated dNTPs and excess primers before multiplex minisequencing was performed. After multiplex minisequencing, unincorporated fluorescent ddNTPs were enzymatically inactivated before the resultant products were resolved by capillary electrophoresis.

As illustrated in Figure 10, this assay was able to unambiguously genotype different DNA samples at the seven MDR1 SNP loci. Each SNP locus was identified based on differences in primer length, and hence migration speeds through the capillary, as well as peak color from incorporation of one of 4 different fluorescent ddNTPs by the SNP-specific primer. For example, the SNP-specific primer of SNP Intron -1/-41A>G is shorter, and thus migrates faster, than the SNP 26 3435C>T primer. In samples homozygous at a particular SNP locus, either of the alternative dye-terminators attached to the SNP-specific primer, producing a single primer peak at that site on the electropherogram. Samples heterozygous for a particular locus had two different dye-terminators attach to the minisequencing primer, producing two,

different-colored peaks. Migration of this same minisequencing primer differed depending on the molecular weight differences of the nucleotide–fluorescent ddNTP combinations. For example, the A-allele peak of SNP exon 21 2677GT>A migrated more slowly than the G-allele peak, so that SNP exon 21 2677GA heterozygous samples displayed two allele peaks with minimal overlap (Figure 10f). However, the T-allele peak did not migrate very differently from the A-allele peak; hence SNP exon 21 2677TA heterozygous sample displayed overlapping allele peaks (Figure 10d). Peak heights of the different alleles could also differ significantly because of differences in the fluorescence intensities of the different fluorophores (Figure 10). Accurate genotypes were obtained from all 100 genotype type-known samples at all seven loci that were tested (data not shown).

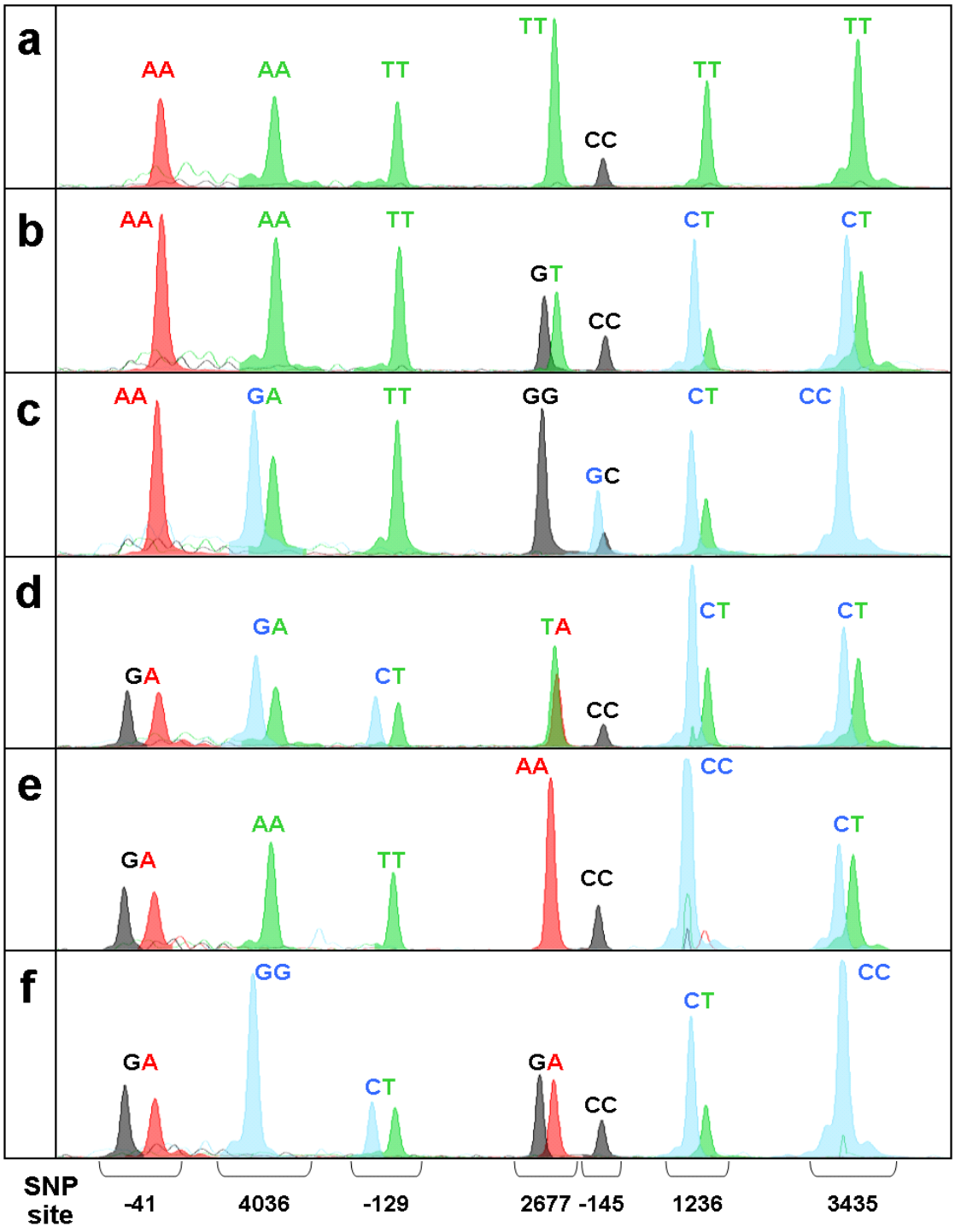


Figure 10. Genotyping results for the seven MDR1 SNPs.
GeneScan 3.7 analysis of multiplex-minisequencing products. Electropherograms of representative samples with different genotypes at the seven SNP sites are shown. Each SNP allele displays a characteristic peak color, position, and height relative to the other allele peak

Chapter 7 Genetic characterization of *ABCC4* gene locus

7.1 Large variability in SNP frequencies amongst populations

Within *ABCC4*, 28 SNPs were genotyped in 4 multiplex minisequencing panels (as shown in Table 8). These SNPs included 12 exonic SNPs, 11 intronic SNPs, 4 SNPs in the 5' flanking region and 1 in the 3' flanking region, all of which were either reported in the public databases or publications (Table 10). SNP i-1 –644C>T and SNP e4 511T>G were monomorphic in all 5 populations while SNP e11 1497C>T was only polymorphic in the African American population. SNP e11 1497C>T may therefore be an ancient variant which was lost after the diversification of the ancient African population into other subpopulations. Only 3 SNPs including the exonic SNP e18 2269G>A possessed a minor allele frequency of less than 5%. The rest of the SNPs were highly polymorphic.

In addition, results also showed allele frequencies to vary widely amongst different populations. Unlike previous studies on MDR1 and *ABCC5*, the Chinese and Malay populations did not share as much similarity in allele frequencies (Tang et al. 2004; Gwee et al. 2005). Previous similar relationship between the Indian and European American populations was also not seen here. However, the African American population remained the most different in terms of allele frequencies. An example could be seen in the C allele of the SNP i19 10304C>A. In the African American population, the C allele occupied an allele frequency of 25.5% and this frequency rose to a range of 66.0-78.5% in the non-African populations.

Positional information of the SNP loci is presented in Table 10 and Figure 8. Ten of 140 different SNP locus/population combinations showed deviations from HWE at α=0.05. Genotyping error was ruled out by rechecking minisequencing electropherograms, sequencing random samples as well as samples in which

genotypes were too difficult to call. These deviations may therefore be due to other violations of assumptions inherent to Hardy-Weinberg such as close linkage to areas under selection. Most of these deviations were also observed to occur in the African American population. Although its ancestry is predominantly African, it may be difficult to rule out admixture from European and Native American founding populations (Parra et al., 1988; Kittles et al., 2002). Upon Bonferroni adjustment for multiple-loci testing, no significant evidence for deviation from Hardy-Weinberg equilibrium (HWE) was found in all SNPs.

SNP no.	SNP ID	dbSNP rs#	JSNP id	TSC SNP id	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's exact P-value)					Fst 5pop	Fst 3pop	Pairwise Fst Value					Pexcess
															CH	ML	IN	CAU	AA			CH	ML	IN	CAU	AA	
1	i-1 -1393G>A	rs868853	-	-	5' Flanking					AA	GG	AG	A	G													
							CH	93	0.56	72.04	3.23	24.73	84.41	15.59		0.23	0.09	0.03	6.56E-03	0.04	0.06		0.01	0.01	0.02	0.03	0.70
							ML	96	0.61	63.54	5.21	31.25	79.17	20.83			2.61E-03	2.86E-04	0.16					0.05	0.06	0.00	
							IN	94	0.30	80.85	0.00	19.15	90.43	9.57				0.60	1.21E-05						0.00	0.09	
							CAU	100	0.62	85.00	1.00	14.00	92.00	8.00					6.70E-07							0.11	
							AA	100	0.51	52.00	6.00	42.00	73.00	27.00													
2	i-1 -1015G>A	rs2993579	-	-	5' Flanking					AA	GG	AG	A	G													
							CH	93	4.04E-03	33.33	6.45	60.22	63.44	36.56		5.16E-03	0.29	0.30	4.83E-12	0.07	0.13		0.04	0.00	0.00	0.21	0.41
							ML	96	0.42	26.04	28.13	45.83	48.96	51.04			0.08	0.09	3.32E-05					0.01	0.01	0.08	
							IN	94	0.27	30.85	14.89	54.26	57.98	42.02				1.00	4.28E-09						0.00	0.16	
							CAU	100	0.79	33.00	17.00	50.00	58.00	42.00					3.66E-09							0.16	
							AA	100	0.13	5.00	48.00	47.00	28.50	71.50													
3	i-1 -644C>T	rs3814270	-	-	5' Flanking					CC	TT	CT	C	T													
							CH	93	1.00	100.00	0.00	0.00	100.00	0.00		1.00	1.00	1.00	1.00								
							ML	96	1.00	100.00	0.00	0.00	100.00	0.00			1.00	1.00	1.00								
							IN	94	1.00	100.00	0.00	0.00	100.00	0.00				1.00	1.00								
							CAU	100	1.00	100.00	0.00	0.00	100.00	0.00					1.00								
							AA	100	1.00	100.00	0.00	0.00	100.00	0.00													
4	i-1 -527C>G	rs869951	-	-	5' Flanking					CC	GG	CG	C	G													
							CH	93	9.70E-03	35.48	6.45	58.06	64.52	35.48		1.31E-03	0.11	0.10	1.09E-12	0.07	0.13		0.05	0.01	0.01	0.23	0.38
							ML	96	0.70	23.96	28.13	47.92	47.92	52.08			0.10	0.13	7.82E-05					0.01	0.01	0.07	
							IN	94	0.43	29.79	17.02	53.19	56.38	43.62				1.00	3.96E-08						0.00	0.14	
							CAU	100	0.88	31.00	19.00	50.00	56.00	44.00					3.73E-08							0.14	
							AA	100	0.13	5.00	48.00	47.00	28.50	71.50													
5	e1 -49C>T	rs3751333	ssj0000449	-	Exon 1					CC	TT	CT	C	T													
			IMS-JST103375				CH	93	0.88	81.72	1.08	17.20	90.32	9.68		0.42	0.02	4.20E-03	1.60E-05	0.04	0.05		0.00	0.02	0.04	0.08	0.02
							ML	96	0.64	76.04	1.04	22.92	87.50	12.50			2.26E-03	1.55E-04	2.56E-07					0.05	0.07	0.11	
							IN	94	0.71	92.55	0.00	7.45	96.28	3.72				0.57	0.03						0.00	0.02	
							CAU	100	0.80	95.00	0.00	5.00	97.50	2.50					0.22							0.01	
							AA	100	0.96	99.00	0.00	1.00	99.50	0.50													
6	i1 17905G>A	rs9524885	-	-	Intron 1					AA	GG	AG	A	G													
							CH	93	0.46	13.98	34.41	51.61	39.78	60.22		2.88E-03	0.11	0.09	4.31E-03	0.05	0.05		0.04	0.01	0.01	0.04	0.44
							ML	96	0.60	29.17	18.75	52.08	55.21	44.79			3.05E-06	1.48E-06	0.92					0.11	0.11	0.00	
							IN	94	0.90	9.57	46.81	43.62	31.38	68.62				1.00	5.89E-06						0.00	0.10	
							CAU	100	0.22	7.00	45.00	48.00	31.00	69.00					2.99E-06							0.10	
							AA	100	0.02	24.00	15.00	61.00	54.50	45.50													
7	i1 34530T>C	rs871175	-	TSC0140541	Intron 1					CC	TT	CT	C	T													
							CH	93	0.10	1.08	62.37	36.56	19.35	80.65		0.18	1.76E-04	6.28E-09	0.32	0.08	0.10		0.01	0.07	0.15	0.00	0.91
							ML	96	0.89	6.25	55.21	38.54	25.52	74.48			2.63E-07	9.73E-13	0.81					0.12	0.21	0.00	
							IN	94	0.51	0.00	87.23	12.77	6.38	93.62				0.04	1.12E-06						0.02	0.10	
							CAU	100	0.84	0.00	96.00	4.00	2.00	98.00					8.51E-12							0.19	
							AA	100	0.90	6.00	58.00	36.00	24.00	76.00													

Table 10. Genotype and allele frequency analysis of 28 SNPs at the *ABCC*4 gene locus.

SNP no.	SNP ID	dbSNP rs#	JSNP id	TSC SNP id	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's exact P-value)					Fst 5pop	Fst 3pop	Pairwise Fst Value					Pexcess
															CH	ML	IN	CAU	AA			CH	ML	IN	CAU	AA	
										CC	TT	CT	C	T													
8	i2 290C>T	rs4148436	-	-	Intron 2		CH	93	0.05	9.68	33.33	56.99	38.17	61.83		0.02	0.03	0.92	0.92	0.03	0.00		0.03	0.03	0.00	0.00	0.02
							ML	96	0.31	28.13	27.08	44.79	50.52	49.48			3.37E-06	0.03	0.01					0.12	0.02	0.03
							IN	94	0.63	6.38	52.13	41.49	27.13	72.87				0.01	0.03						0.03	0.02
							CAU	100	0.45	17.00	39.00	44.00	39.00	61.00					0.84							0.00
							AA	100	0.38	12.00	37.00	51.00	37.50	62.50												
										TT	GG	TG	T	G													
9	e4 511T>G	rs4148460	ssj0000492	-	Exon 4	Cys171Gly	CH	93	1.00	100.00	0.00	0.00	100.00	0.00	1.00	1.00	1.00	1.00									
							ML	96	1.00	100.00	0.00	0.00	100.00	0.00		1.00	1.00	1.00									
							IN	94	1.00	100.00	0.00	0.00	100.00	0.00			1.00	1.00									
							CAU	100	1.00	100.00	0.00	0.00	100.00	0.00				1.00									
							AA	100	1.00	100.00	0.00	0.00	100.00	0.00													
										CC	TT	CT	C	T													
10	i4 5124T>C	rs2389234	-	TSC1590442	Intron 4		CH	93	0.67	7.53	49.46	43.01	29.03	70.97	0.10	1.00	0.02	0.19		0.02	0.03		0.01	0.00	0.02	0.00	0.47
							ML	96	0.11	7.29	64.58	28.13	21.35	78.65		0.08	0.53	2.46E-03						0.02	0.00	0.05
							IN	94	0.13	5.32	46.81	47.87	29.26	70.74			0.02	0.20							0.03	0.00
							CAU	100	0.35	2.00	65.00	33.00	18.50	81.50				1.86E-04								0.06
							AA	100	0.48	11.00	40.00	49.00	35.50	64.50												
										CC	TT	CT	C	T													
11	e6 669C>T	rs899494	ssj0000521	TSC0174780	Exon 6	Ile223Ile	CH	93	0.38	64.52	2.15	33.33	81.18	18.82	0.69	0.61	0.69	0.38		0.00	0.00		0.00	0.00	0.00	0.00	0.24
			IMS-JST179043				ML	96	0.12	70.83	5.21	23.96	82.81	17.19		0.36	1.00	0.21						0.00	0.00	0.01
							IN	94	0.88	61.70	4.26	34.04	78.72	21.28			0.30	0.81							0.00	0.00
							CAU	100	0.18	67.00	1.00	32.00	83.00	17.00				0.21								0.00
							AA	100	0.02	56.00	1.00	43.00	77.50	22.50												
										GG	TT	GT	G	T													
12	e8 912G>T	rs2274407	ssj0000529	-	Exon 8	Lys302Asn	CH	93	0.76	69.89	3.23	26.88	83.33	16.67	0.69	0.79	0.03	0.48		0.01	0.01		0.00	0.00	0.02	0.00	0.36
			IMS-JST070256				ML	96	0.68	66.67	4.17	29.17	81.25	18.75		0.90	5.41E-03	0.22						0.00	0.03	0.00
							IN	94	0.45	65.96	2.13	31.91	81.91	18.09			0.01	0.33							0.03	0.00
							CAU	100	0.15	84.00	2.00	14.00	91.00	9.00				0.16								0.01
							AA	100	0.43	73.00	1.00	26.00	86.00	14.00												
										AA	GG	AG	A	G													
13	e8 951G>A	rs2274406	ssj0000530	-	Exon 8	Arg317Arg	CH	93	0.51	22.58	31.18	46.24	45.70	54.30	0.12	0.03	0.10	1.41E-05		0.07	0.10		0.01	0.02	0.01	0.09	0.46
			IMS-JST070255				ML	96	0.33	31.25	23.96	44.79	53.65	46.35		2.01E-04	1.14E-03	3.82E-03						0.06	0.05	0.04
							IN	94	0.14	8.51	39.36	52.13	34.57	65.43			0.67	5.23E-11							0.00	0.20
							CAU	100	0.47	12.00	38.00	50.00	37.00	63.00				7.46E-10								0.18
							AA	100	0.91	46.00	10.00	44.00	68.00	32.00												
										AA	GG	AG	A	G													
14	e8 969G>A	rs2274405	ssj0000531	-	Exon 8	Ser323Ser	CH	93	0.51	22.58	31.18	46.24	45.70	54.30	0.12	0.01	0.06	0.06		0.02	0.01		0.01	0.03	0.01	0.01	0.00
			IMS-JST070254				ML	96	0.33	31.25	23.96	44.79	53.65	46.35		3.36E-05	5.40E-04	5.40E-04						0.08	0.05	0.05
							IN	94	0.17	7.45	42.55	50.00	32.45	67.55			0.52	0.52							0.00	0.00
							CAU	100	0.68	12.00	40.00	48.00	36.00	64.00				1.00								0.00
							AA	100	0.39	11.00	39.00	50.00	36.00	64.00												

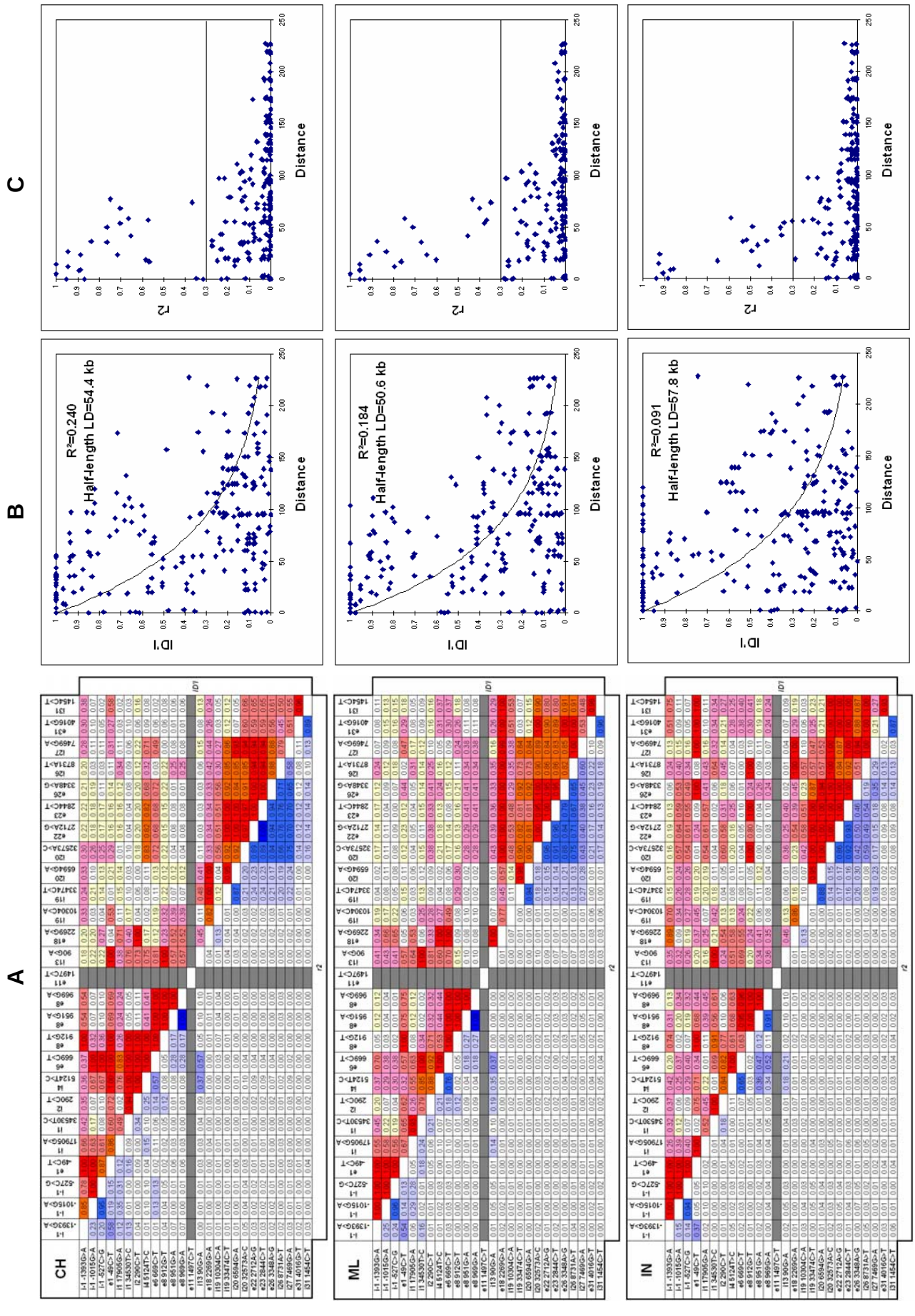
Table 10 (continued).

SNP no.	SNP ID	dbSNP rs#	JSNP id	TSC SNP id	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's exact P-value)					Fst 5pop	Fst 3pop	Pairwise Fst Value					Pexcess
															CH	ML	IN	CAU	AA			CH	ML	IN	CAU	AA	
15	e11 1497C>T	rs1557070	-	TSC0455420	Exon 11	Tyr499Tyr				CC	TT	CT	C	T												
							CH	93	1.00	100.00	0.00	0.00	100.00	0.00		1.00	1.00	1.00	8.19E-15	0.25	0.25					0.24	1.00
							ML	96	1.00	100.00	0.00	0.00	100.00	0.00			1.00	1.00	1.03E-14							0.25	
							IN	94	1.00	100.00	0.00	0.00	100.00	0.00				1.00	8.15E-15							0.24	
							CAU	100	1.00	100.00	0.00	0.00	100.00	0.00					1.58E-15							0.25	
							AA	100	0.43	54.00	5.00	41.00	74.50	25.50													
16	i13 90G>A	rs1751005	ssj0000564	TSC0044842	Intron 13					AA	GG	AG	A	G												
			IMS-JST050915				CH	93	0.50	3.23	61.29	35.48	20.97	79.03	0.90	4.20E-04	0.47	8.02E-03	0.05	0.03	0.00	0.06	0.00	0.03	0.15	
							ML	96	0.12	2.08	58.33	39.58	21.88	78.13		7.61E-04	0.55	4.03E-03				0.06	0.00	0.03		
							IN	94	0.54	12.77	37.23	50.00	37.77	62.23			5.92E-03	5.87E-10					0.03	0.17		
							CAU	100	6.90E-03	11.00	62.00	27.00	24.50	75.50				5.98E-04						0.06		
							AA	100	0.22	0.00	78.00	22.00	11.00	89.00													
17	e18 2269G>A	rs3765534	ssj0000585	-	Exon 18	Glu757Lys				AA	GG	AG	A	G												
			IMS-JST119723				CH	93	0.62	0.00	90.32	9.68	4.84	95.16	0.62	0.21	0.08	8.55E-03	0.02	0.02	0.00	0.01	0.01	0.03	0.01	
							ML	96	0.71	0.00	92.71	7.29	3.65	96.35		0.05	0.21	0.03				0.02	0.00	0.02		
							IN	94	0.37	0.00	82.98	17.02	8.51	91.49			1.59E-03	7.20E-05					0.05	0.07		
							CAU	100	0.88	0.00	97.00	3.00	1.50	98.50				0.62						0.00		
							AA	100	0.96	0.00	99.00	1.00	0.50	99.50													
18	19 10304C>A	rs1479390	ssj0000592	TSC0703180	Intron 19					AA	CC	AC	A	C												
							CH	93	0.67	5.38	62.37	32.26	21.51	78.49	0.71	7.90E-03	6.60E-03	9.83E-15	0.19	0.28	0.00	0.04	0.03	0.44	0.54	
							ML	96	0.88	5.21	58.33	36.46	23.44	76.56		0.02	0.03	1.66E-14				0.02	0.02	0.40		
							IN	94	0.18	8.51	40.43	51.06	34.04	65.96			1.00	8.80E-15					0.00	0.27		
							CAU	100	0.49	10.00	42.00	48.00	34.00	66.00				1.10E-14						0.28		
							AA	100	0.19	58.00	9.00	33.00	74.50	25.50													
19	19 33474C>T	rs1189429	ssj0000607	TSC1769539	Intron 19					CC	TT	CT	C	T												
							CH	93	0.32	31.18	15.05	53.76	58.06	41.94	1.39E-03	4.52E-06	0.10	0.17	0.06	0.02	0.05	0.11	0.01	0.00	0.23	
							ML	96	0.60	15.63	33.33	51.04	41.15	58.85		0.17	0.11	2.91E-06				0.01	0.01	0.10		
							IN	94	0.33	13.83	45.74	40.43	34.04	65.96			2.79E-03	1.54E-09					0.04	0.17		
							CAU	100	0.55	23.00	24.00	53.00	49.50	50.50				2.38E-03						0.04		
							AA	100	0.44	44.00	14.00	42.00	65.00	35.00													
20	i20 6594G>A	rs2766481	-	-	Intron 20					AA	GG	AG	A	G												
			IMS-JST139428				CH	93	0.84	17.20	35.48	47.31	40.86	59.14	4.54E-04	9.58E-07	0.08	2.87E-07	0.14	0.12	0.06	0.12	0.01	0.12	0.40	
							ML	96	0.44	33.33	14.58	52.08	59.38	40.63		0.17	0.07	1.38E-14				0.01	0.01	0.31		
							IN	94	0.50	45.74	12.77	41.49	66.49	33.51			1.39E-03	2.04E-14					0.05	0.40		
							CAU	100	0.69	24.00	24.00	52.00	50.00	50.00				2.68E-12						0.22		
							AA	100	0.13	5.00	71.00	24.00	17.00	83.00													
21	20 32573A>C	rs1189437	ssj0000643	-	Intron 20					AA	CC	AC	A	C												
							CH	93	0.92	54.84	6.45	38.71	74.19	25.81	1.76E-03	5.86E-07	9.74E-07	0.81	0.06	0.06	0.05	0.12	0.11	0.00	0.69	
							ML	96	0.14	73.96	0.00	26.04	86.98	13.02		0.06	0.09	4.37E-03				0.02	0.01	0.04		
							IN	94	0.38	87.23	1.06	11.70	93.09	6.91			0.85	1.66E-06					0.00	0.11		
							CAU	100	0.53	86.00	1.00	13.00	92.50	7.50				4.43E-06						0.10		
							AA	100	6.49E-03	67.00	16.00	17.00	75.50	24.50													

Table 10 (continued).

SNP no.	SNP ID	dbSNP rs#	JSNP id	TSC SNP id	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's exact P-value)					Fst 5pop	Fst 3pop	Pairwise Fst Value					Pexcess
															CH	ML	IN	CAU	AA			CH	ML	IN	CAU	AA	
										AA	GG	AG	A	G													
22	e22 2712A>G	rs1678339	ssj0000655 IMS-JST070250	-	Exon 22	Leu904Leu	CH	93	0.86	6.45	56.99	36.56	24.73	75.27		6.01E-03	1.09E-05	2.07E-06	1.43E-03	0.04	0.06		0.04	0.09	0.11	0.05	0.42
							ML	96	0.12	0.00	72.92	27.08	13.54	86.46			0.10	0.04	0.65					0.01	0.02	0.00	
							IN	94	0.57	1.06	85.11	13.83	7.98	92.02				0.85	0.24						0.00	0.00	
							CAU	100	0.45	0.00	86.00	14.00	7.00	93.00					0.12							0.01	
							AA	100	0.60	2.00	78.00	20.00	12.00	88.00													
										CC	TT	CT	C	T													
23	e23 2844C>T	rs1189466	ssj0000657 IMS-JST070249	-	Exon 23	Phe948Phe	CH	93	0.86	56.99	6.45	36.56	75.27	24.73		3.80E-03	4.64E-06	7.93E-07	1.58E-05	0.05	0.08		0.04	0.10	0.12	0.09	0.24
							ML	96	0.14	73.96	0.00	26.04	86.98	13.02			0.09	0.04	0.19					0.01	0.02	0.01	
							IN	94	0.47	86.17	1.06	12.77	92.55	7.45				0.84	0.71						0.00	0.00	
							CAU	100	0.49	87.00	0.00	13.00	93.50	6.50					0.57							0.00	
							AA	100	0.72	84.00	1.00	15.00	91.50	8.50													
										AA	GG	AG	A	G													
24	e26 3348A>G	rs1751034	ssj0000672	-	Exon 26	Lys1116Lys	CH	93	0.66	55.91	7.53	36.56	74.19	25.81		7.08E-03	0.01	0.09	0.65	0.02	0.01		0.03	0.03	0.01	0.00	0.34
							ML	96	0.39	71.88	1.04	27.08	85.42	14.58			1.00	0.34	1.33E-03					0.00	0.00	0.05	
							IN	94	0.94	72.34	2.13	25.53	85.11	14.89				0.41	1.99E-03						0.00	0.05	
							CAU	100	0.11	64.00	1.00	35.00	81.50	18.50					0.03							0.02	
							AA	100	0.01	57.00	13.00	30.00	72.00	28.00													
										AA	TT	AT	A	T													
25	i26 8731A>T	rs1211485	-	TSC1590319	Intron 26		CH	93	0.61	51.61	6.45	41.94	72.58	27.42		1.46E-03	7.28E-05	1.12E-05	0.02	0.03	0.04		0.05	0.08	0.09	0.02	0.43
							ML	96	0.11	71.88	0.00	28.13	85.94	14.06			0.44	0.28	0.41					0.00	0.00	0.00	
							IN	94	0.86	78.72	1.06	20.21	88.83	11.17				0.74	0.08						0.00	0.01	
							CAU	100	0.27	80.00	0.00	20.00	90.00	10.00					0.04							0.02	
							AA	100	0.18	70.00	5.00	25.00	82.50	17.50													
										AA	GG	AG	A	G													
26	i27 7469G>A	rs1151471	ssj0000683	-	Intron 27		CH	93	0.67	49.46	7.53	43.01	70.97	29.03	0.10	0.82	0.05	2.37E-08	0.08	0.08		0.01	0.00	0.02	0.15	0.32
							ML	96	0.40	60.42	3.13	36.46	78.65	21.35		0.05	2.64E-04	2.42E-13					0.01	0.06	0.23	
							IN	94	0.20	45.74	6.38	47.87	69.68	30.32			0.11	8.89E-08						0.01	0.14	
							CAU	100	0.23	35.00	12.00	53.00	61.50	38.50				2.05E-04							0.06	
							AA	100	0.43	20.00	35.00	45.00	42.50	57.50													
										GG	TT	GT	G	T													
27	e31 4016G>T	rs3742106	ssj0000696 IMS-JST093204	-	Exon 31		CH	93	0.25	23.66	20.43	55.91	51.61	48.39	0.26	0.84	1.03E-03	4.39E-04	0.04	0.04		0.00	0.00	0.05	0.07	0.02
							ML	96	0.65	32.29	16.67	51.04	57.81	42.19		0.18	7.64E-06	1.68E-06					0.00	0.10	0.11	
							IN	94	0.22	22.34	21.28	56.38	50.53	49.47			2.11E-03	9.54E-04						0.05	0.06	
							CAU	100	0.23	15.00	45.00	40.00	35.00	65.00				0.83							0.00	
							AA	100	0.03	16.00	49.00	35.00	33.50	66.50													
										CC	TT	CT	C	T													
28	i31 1454C>T	rs1059762	-	-	Intron 31		CH	93	0.35	21.51	23.66	54.84	48.92	51.08	0.22	0.92	5.69E-03	0.07	0.02	0.02		0.00	0.00	0.04	0.02	0.11
							ML	96	0.97	17.71	33.33	48.96	42.19	57.81		0.18	4.90E-05	1.70E-03					0.00	0.08	0.05	
							IN	94	0.01	18.09	19.15	62.77	49.47	50.53			7.85E-03	0.08						0.04	0.01	
							CAU	100	0.32	42.00	16.00	42.00	63.00	37.00				0.41							0.00	
							AA	100	0.02	40.00	23.00	37.00	58.50	41.50													

Table 10 (continued).


A
B
C
I
II
III
R²=0.240
Half-length LD=54.4 kb
R²=0.184
Half-length LD=50.6 kb
R²=0.091
Half-length LD=57.8 kb
Distance
r2
CH
ML
IN

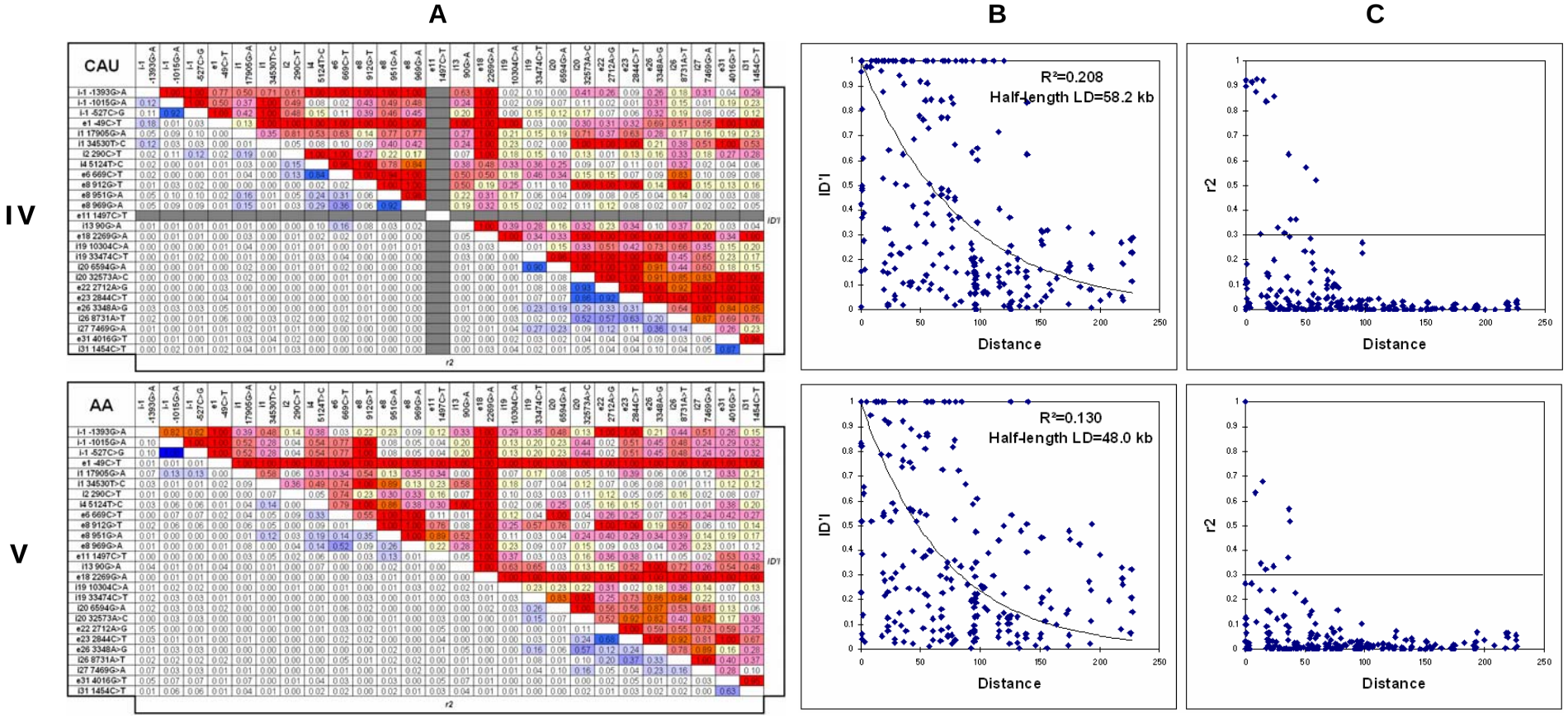


Figure 11. Pair-wise linkage disequilibrium profiles of *ABCC4* gene locus.
Panel A: |*D'*| and r^2 values for each pair of the 25 highly polymorphic SNPs in five populations were summarized into tables [I-Chinese (CH); II-Malay (ML); III-Indian (IN); IV-European American (CAU); V-African American (AA)]. Cells in the triangular section above the diagonal indicate |*D'*| values. Cells in the triangular section below the diagonal contain the r^2 results. Gradient of either red (|*D'*|) or blue (r^2) color for each cell is presented according to the scale specified beside the tables and graphs, with colors of greater intensity representing higher values of |*D'*| or r^2. "-" indicates pairs where calculations of LD cannot be applied. Panel B: |*D'*| between pairs of single SNPs was calculated (see Methods) and plotted against physical distance in the five populations. The correlation coefficient R^2, as well as the half-length LD distances are shown next to the fitted line. Panel C: r^2 between pairs of single SNPs was calculated (see Methods) and plotted against physical distance in the five populations. The threshold for useful LD is set at $r^2 \geq 0.3$.

7.2 Rapidly declining LD profile across *ABCC4* gene locus

The LD between all pairs of genotyped SNP loci was investigated and the data presented in Figure 11. Two traditional measures of LD, *D'* and r^2, were used to estimate recombination rates and strength of association respectively. Pair-wise LD in Figure 11 showed that general pattern of LD across all 5 populations was similar. SNPs in close proximity showed more intense LD between them than those apart e.g. SNPs e22 2712A>G and e23 2844C>T. SNP e18 2269G>A showed erroneously high measures of LD with all other SNPs in the European American and African American populations, due to its low minor allele frequency of less than 2%. Since SNP e11 1497C>T only occurred in the African American population, LD values were not calculated for the rest of the populations (highlighted in grey in Figure 11).

LD measures against genomic distance were correlated, restricting values from SNPs with minor allele frequency >5%. The common LD decay pattern existed in all populations, although the negative correlation between LD measures and distance was poor (Figure 11A). Half-length LD of the Chinese, Malay, Indian, European American and African American populations were 54.4 kb, 50.6kb, 57.8kb, 58.2 kb and 48.0kb, with correlation coefficient R^2 at 0.240, 0.184, 0.091, 0.208, 0.130 respectively (Figure 11B). In addition, useful r^2 values greater than 0.3 were estimated to be 77.42 kb in both Chinese and Malays, 59.03 kb in both Indians and European Americans, and 37.23 kb in the African Americans (Figure 11C).

7.3 High variability in profiles of common haplotypes amongst populations

The overall LD across the *ABCC4* gene was rather weak. Consequently, the number of haplotypes for the gene was likely to be large. Only a short stretch between SNPs e18 2269G>A and i27 7469G>A showed discernibly tight linkage in all populations. The number of haplotypes was inferred, based on estimates derived using the EM algorithm from the genotypic data of 25 polymorphic SNPs in all 5 populations.

EM estimated that there were 104, 100, 116, 117, and 151 haplotypes in the Chinese, Malay, Indian, European American and African American populations respectively. In total, 536 unique haplotypes could be found for all samples. None of the estimated haplotypes was shared by all 5 populations and only 10 haplotypes were shared by more than 2 populations. These numbers of EM-estimated haplotypes were compared to those derived from a simulated population of the same sample size and observed allele frequency assuming random association among the 25 analyzed SNPs. Under random association, 185.95±0.23, 191.88±0.34, 187.61±0.62, 199.31±0.86, and 199.68±0.57 haplotypes were predicted to exist in the Chinese, Malay, Indian, European American and African American populations respectively. The reduction in predicted haplotype number (~ 52.1 - 75.6%) represented a small but significant amount of recombination that has taken place in this locus.

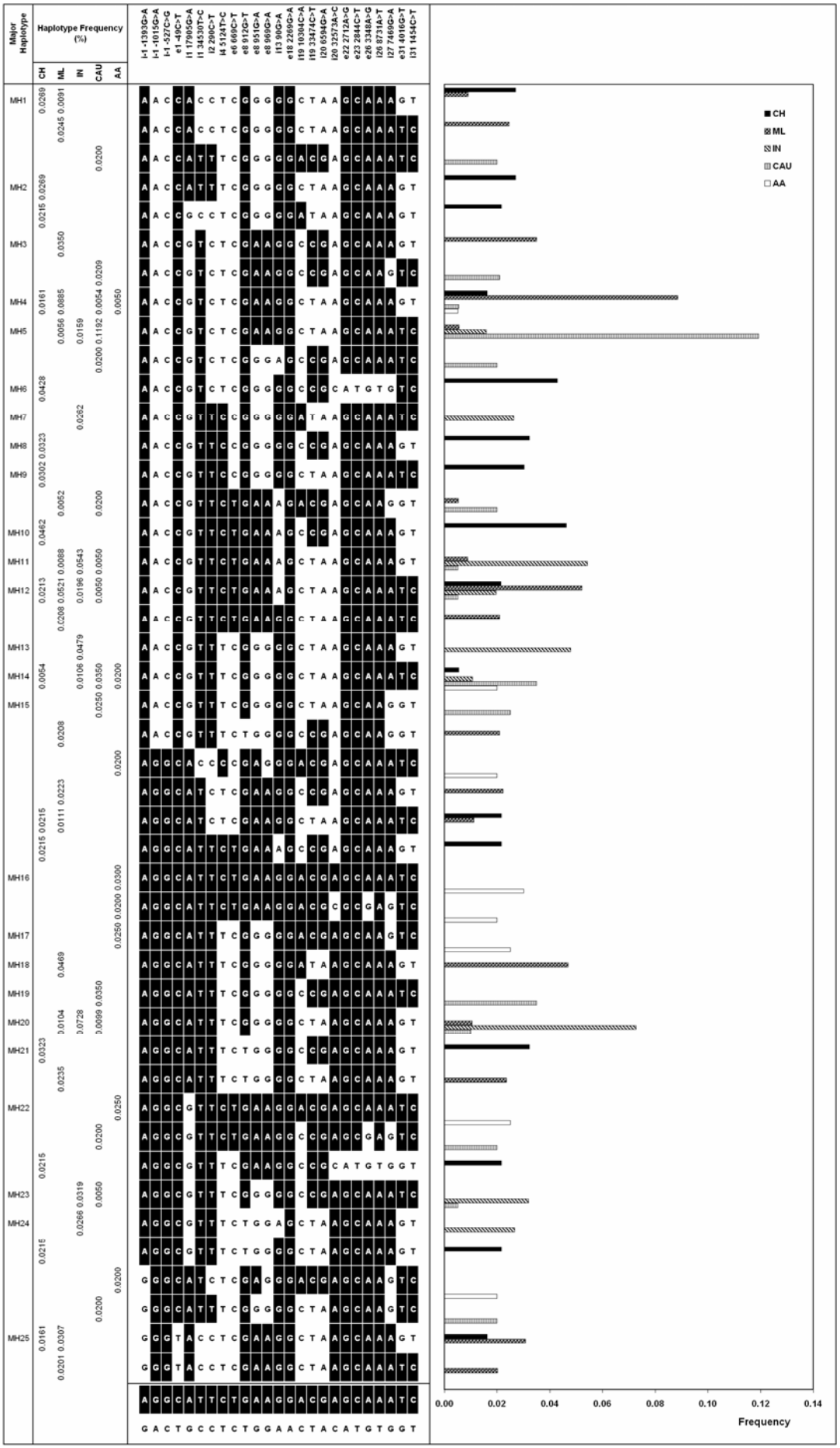

Major Haplotype
Haplotype Frequency (%)
CH
ML
IN
CAU
AA
I-1 -1393G>A
I-1 -1015G>A
I-1 -527C>G
e1 -49C>T
I1 17905G>A
I1 34530T>C
I2 290C>T
I4 5124T>C
e6 669C>T
e8 912G>T
e8 951G>A
e8 969G>A
I13 90G>A
e18 2269G>A
I19 10304C>A
I19 33474C>T
I20 6594G>A
I20 32573A>C
e22 2712A>G
e23 2844C>T
e26 33468A>G
I26 8731A>T
I27 7469G>A
e31 4016G>T
I31 1454C>T
MH1
MH2
MH3
MH4
MH5
MH6
MH7
MH8
MH9
MH10
MH11
MH12
MH13
MH14
MH15
MH16
MH17
MH18
MH19
MH20
MH21
MH22
MH23
MH24
MH25
G A C T G C C T C T G G A A C T A C A T G T G G T
CH
ML
IN
CAU
AA
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Frequency

Figure 12. Haplotype profiles of SNPs at the *ABCC*4 gene locus. Haplotype frequencies were inferred from genotype data of the 25 highly polymorphic SNPs at the *ABCC*4 gene locus using Expectation-Maximization algorithm. A total of 536 unique haplotypes occur in at least one population, of which 104, 100, 116, 117, and 151 were found in the Chinese, Malay, Indian, European American and African American populations, respectively. Only haplotypes occurring at >2% are presented in this figure and those with frequencies >2.5% are labeled as MH1-MH25. Haplotype profiles are represented as horizontal arrays of black or white boxes with the allele specified. Each column of boxes represents a SNP locus. Each row of boxes represents a haplotype, with the estimated frequencies expressed as a percentage shown in the adjacent table as well as portrayed in a bar graph. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American).

In Figure 12, 25 haplotypes with frequency of at least 2.5% in any 1 population are identified as "major haplotypes" (MH). The highest haplotype frequencies observed are 4.62% (MH10 in Chinese), 8.85% (MH4 in Malays), 7.28% (MH20 in Indians), 11.92% (MH5 in European Americans) and 3.00% (MH15 in African Americans).

Since the haplotype distribution profiles of the populations differed widely, it would be important to find out if the differences observed were significant. Applying Fisher's Exact test to the haplotype frequencies showed that the African American population was significantly different from the rest of the 5 populations ($p<0.05$), with the exception of the Indian population. No significant differences could be observed in other pair-wise population differentiation.

7.4 Population-specific tagging SNP sets

The presence of low-grade LD across the *ABCC*4 gene signified that there were considerably large sets of haplotypes. However this study still attempted to reduce the complexity of the haplotypes by calculating the smallest tagging-SNP (tSNPs) sets sufficient to represent all the haplotypes for each population (Table 11). A total of at least 7 tSNPs was required to obtain the weighted-average haplotype r^2 values of at least 90% of maximum value in non-African populations whereas a minimum of 9 tSNPs was required for African Americans. Table 11 shows that the identity of the SNPs which may be most suitable for use as tSNPs for the association studies. None of the sets were the same for any 2 populations, which was an indication that specific sets of tSNPs need to be discovered for individual populations studied for the use of association studies. However, it also calculated that 10 tSNPs would be enough to obtain 90% coverage if the chromosomes in all of the 5 populations were studied.

number of tSNPs	number of possible tSNP sets	Populations	i-1 -1393G>A	i-1 -1015G>A	i-1 -527C>G	e1 -49C>T	i1 17905G>A	i1 34530T>C	i2 290C>T	i4 5124T>C	e6 669C>T	e8 912G>T	e8 951G>A	e8 969G>A	i13 90G>A	e18 2269G>A	i19 10304C>A	i19 33474C>T	i20 6594G>A	i20 32573A>T	e22 2712A>G	e23 2844C>T	e26 3348A>G	i26 8731A>T	i27 7469G>A	e31 4016G>T	i31 1454C>T	Coverage
H=6	177100	CH		■					■				■	■					■	■						■		0.88365
		ML		■					■				■	■				■						■		■		0.84056
		IN			■				■						■		■	■									■	0.81913
		CAU		■					■					■				■							■		■	0.82871
		AA		■	■		■		■					■									■			■		0.69116
H=7	480700	CH		■					■				■	■			■	■		■						■		0.9366
		ML		■				■					■	■	■				■						■	■	■	0.90665
		IN			■		■						■		■		■	■									■	0.90159
		CAU			■				■					■			■	■					■				■	0.90911
		AA	■	■			■		■					■						■							■	0.79806
H=8	1081575	CH		■					■	■			■				■	■		■						■		0.95847
		ML		■				■		■			■	■			■	■				■				■		0.94894
		IN			■		■		■				■		■		■		■								■	0.95638
		CAU		■			■		■				■				■	■							■	■		0.94767
		AA	■				■		■	■							■			■					■		■	0.88311
		ALL		■			■		■					■			■	■							■		■	0.81991
H=9	2042975	AA		■			■		■	■			■					■		■					■	■		0.9348
		ALL		■			■		■					■			■	■			■				■		■	0.87956
H=10	3268760	ALL	■	■			■		■					■			■	■		■					■		■	0.9224

Table 11. Combinations of tagging SNPs (tSNPs) at the *ABCC*4 gene locus. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American; ALL: all 5 populations).

7.5 Scanning for evidence of positive selection by F_{st} and Pexcess

To assess the degree of population structure present among the five studied populations, the traditional index of allele frequency differentiation F_{st} was first applied to each SNP locus (Table 10 and Figure 13). F_{st} values ranges from 0 (no differentiation amongst populations) to unity (complete differentiation amongst populations). The use of the statistic *P*excess was further explored. As explained in Bersaglieri et al 2004, the rationale of using *P*excess was to detect a marker within particularly long haplotype that rapidly rose in frequency. The change in allele frequency from **P***reference*, prior to the event of selection to **P***test*, during selection could therefore be a measure of selection. To observe the distribution of F_{st} and *P*excess values around the *ABCC4* genomic region, these sets of values were plotted against genomic distance across *ABCC4* using the genotype data from 25 polymorphic SNP markers (Figure 13).

As can be seen from the graph in Figure 13, there was no obvious region within the *ABCC4* gene that showed a consistent elevation of F_{st} and *P*excess values across multiple markers. Two different sets of F_{st}, one derived from 3 of the 5 populations (African American, European American and Chinese) (F_{st}3pop) and the other derived from 5 populations (F_{st}5pop) were compared. The 2 sets of F_{st} did not seem to show any difference – peaks and troughs of each set mirrored each other closely. The mean global F_{st} values for 25 polymorphic SNPs were 0.061±0.061 for 3 populations and 0.052±0.041 for 5 populations. Several SNPs e.g. i-1 -1015G>A (F_{st}3pop=0.13), i-1 -527C>G (F_{st}3pop=0.13), i19 10304C>A (F_{st}3pop=0.28) and i20 6594G>A (F_{st}3pop=0.12) were observed to have relatively higher F_{st} values. These F_{st} values at these SNPs might be indicative of weak positive selection driving the skewness in allele frequencies in certain individual populations. The pair-wise F_{st} values between

populations for all polymorphic SNPs were also calculated to elicit population-specific differentiation (Table 10). The African American population consistently showed higher F_{st} values with other populations at SNPs i-1 -1015G>A, i-1 -527C>G, i19 10304C>A, i20 6594G>A and i27 7469G>A. At certain pairings e.g. between African Americans and Chinese or Malays at SNP i19 10304C>A or between African Americans and Indians at SNP i20 6594G>A, F_{st} values were greater than 0.40 although these were not significantly different from genome-wide F_{st} values (0.123±0.131) especially after multiple-test correction for P-values (Akey et al. 2002). The average *P*excess value for 25 polymorphic SNPs was 0.333±0.236. SNPs with high F_{st} values did not necessarily correspond to SNPs with high *P*excess values and the reverse was true. The sporadically distributed markers that showed unusually high *P*excess values were SNPs i-1 -1393G>A (0.70), i1 34530T>C (0.92), and i20 32573T>C (0.69). There were 7 SNP loci with above-average F_{st} and *P*excess values: i-1 -1015G>A, i-1 -527C>G, i1 34530T>C, e8 951G>A, i19 10304C>A, i20 6594G>A and i20 32573A>C. They represented candidate loci for assessment of recent positive selection using a more powerful long-range haplotype method.

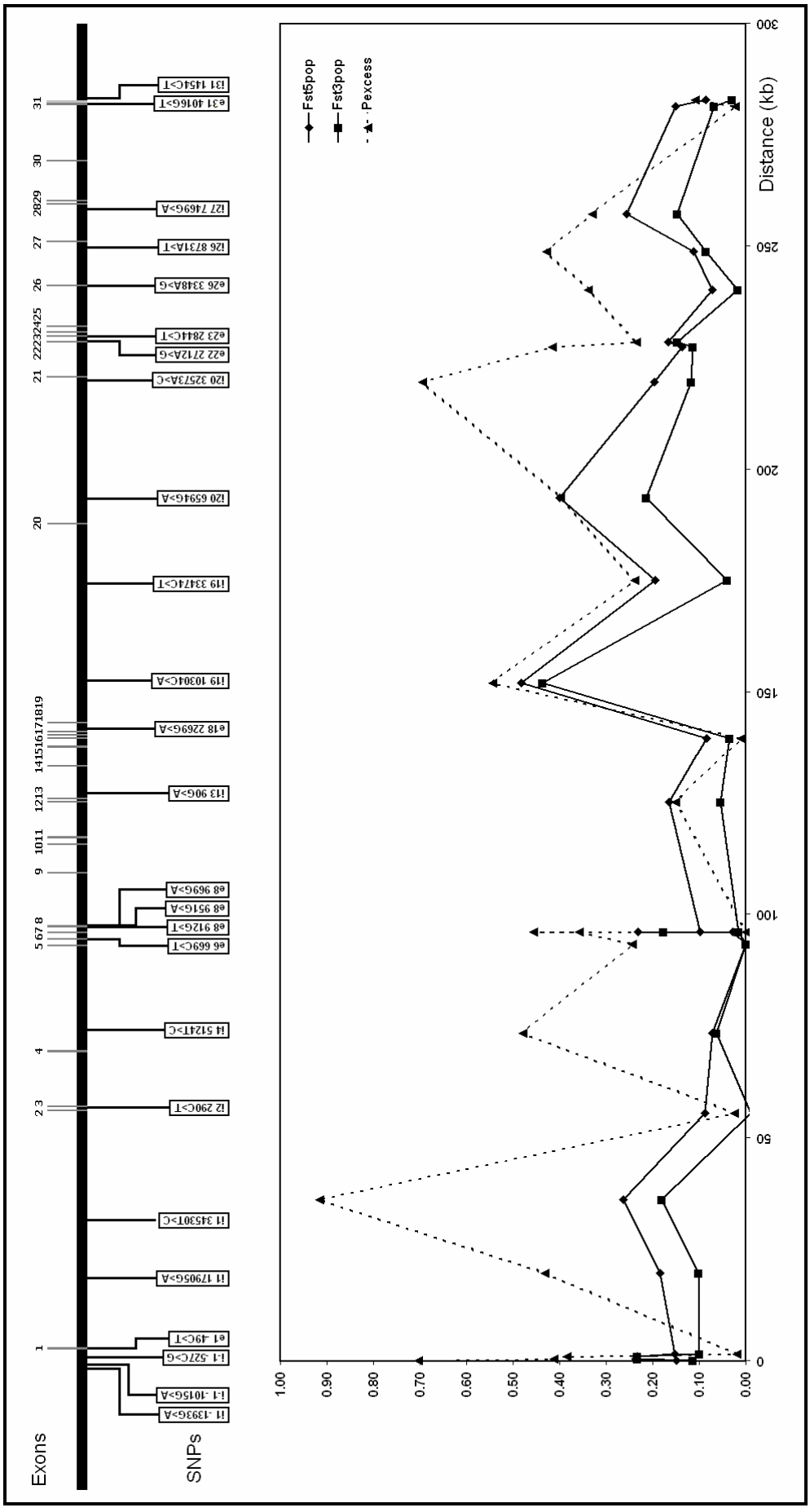


Figure 13. Local profiles of F_{st} and Pexcess values around *ABCC4* gene locus. Fst and Pexcess values for 25 polymorphic SNPs were plotted against genomic distance across the *ABCC4* gene locus. Fst values were calculated for all 5 populations studied (F_{st}5pop) as well as 3

7.6 Revelation of recent positive selection at 5' flanking region by long-range haplotype method

With 125 different SNP locus/population combinations, this study sought to reduce the number of test loci for the long-range haplotype method LRH test. To achieve enough power to test for positive selection, SNP loci with above-average F_{st} and *P*excess values as well as those core alleles with high relative EHH for the longest distance possible were pre-selected for analysis. HBDs were derived, using each of these selected polymorphic SNPs as the core locus to check for evidence of a long-ranging haplotype (Figure 14). Out of all the diagrams derived, four HBDs using SNPs i-1 -1393G>A, i-1 -1015G>A, i-1 -527C>G, and i31 1454C>T as core loci, were selected to be shown here. Diagrams with core SNPs i-1 -1393G>A and i31 1454C>T represented the proximal and distal ends of the *ABCC*4 gene while diagrams with core SNPs i-1 -1015G>A, and i-1 -527C>G showed potential positive selection as shown in Figure 13. Compared with other four ethnic groups, the African Americans showed more number of branches and slightly thinner branches as a result of their rapid decay. As can be seen in HBDs (Figure 14), the major alleles of the markers i-1 -1015G>A, i-1 -527C>G, and i31 1454C>T in the European American population seemed to reside in a long-ranging haplotype across the *ABCC*4 gene locus whereas the minor alleles are on a short-ranging haplotype. This was not observed in other populations, especially that of the African American population. Utilizing a modified long-range haplotype method proposed by Sabeti et al, the relative EHH were plotted against their respective allele frequencies to test for evidence of recent positive selection at *ABCC*4 gene locus (Figure 15). P-values were calculated by comparing relative EHH values against the simulated data at the observed allele frequencies. Based on the HBDs, the core alleles in i-1 -1393G>A, i-1 -1015G>A, i-1

-527C>G, e8 951G>A and i31 1454C>T were selected, all with high relative EHH covering certain distance. The distances from these SNP alleles to the furthest SNP were analyzed e.g. from SNPs located in the 5- region to the last SNP i31 1454C>T.

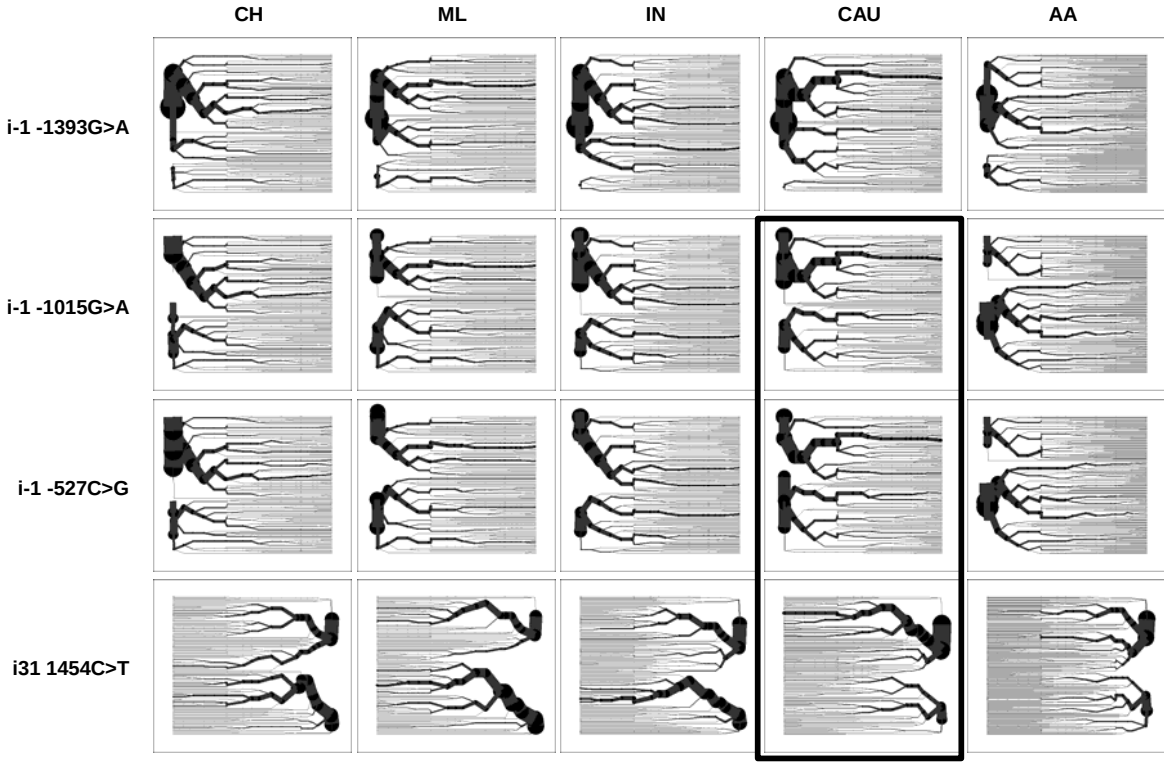


Figure 14. Haplotype Branching Diagrams at 4 selected SNP loci. Haplotype branching diagrams with SNPs i-1 -1393G>A, i-1 -1015G>A, i-1 -527C>G, and i31 1454C>T selected as the roots or core marker locus are presented. SNPs i-1 -1393G>A and i31 1454C>T were selected as the 5' and 3' ends of the *ABCC*4 gene locus. HBDs using each root locus for each population are arranged horizontally. Each root or core marker locus is denoted by a black dot. The position of each locus along the X-axis is scaled to the physical distance between each locus. Each SNP locus upstream or downstream of the root locus is denoted as a node specified by a black dot. Grey lines are drawn between nodes to represent the specific haplotypes. LD decay is portrayed progressively with increasing distance away from the test locus. Branching depends on the presence of alternative alleles at the next consecutive node. The thickness of the branches corresponds to haplotype frequency while the size of the black dots corresponds to allele frequency. HBDs of these SNPs i-1 -1015G>A, i-1 -527C>G, and i31 1454C>T in the European American population are potential candidates for the LRH tests and are therefore highlighted. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American).

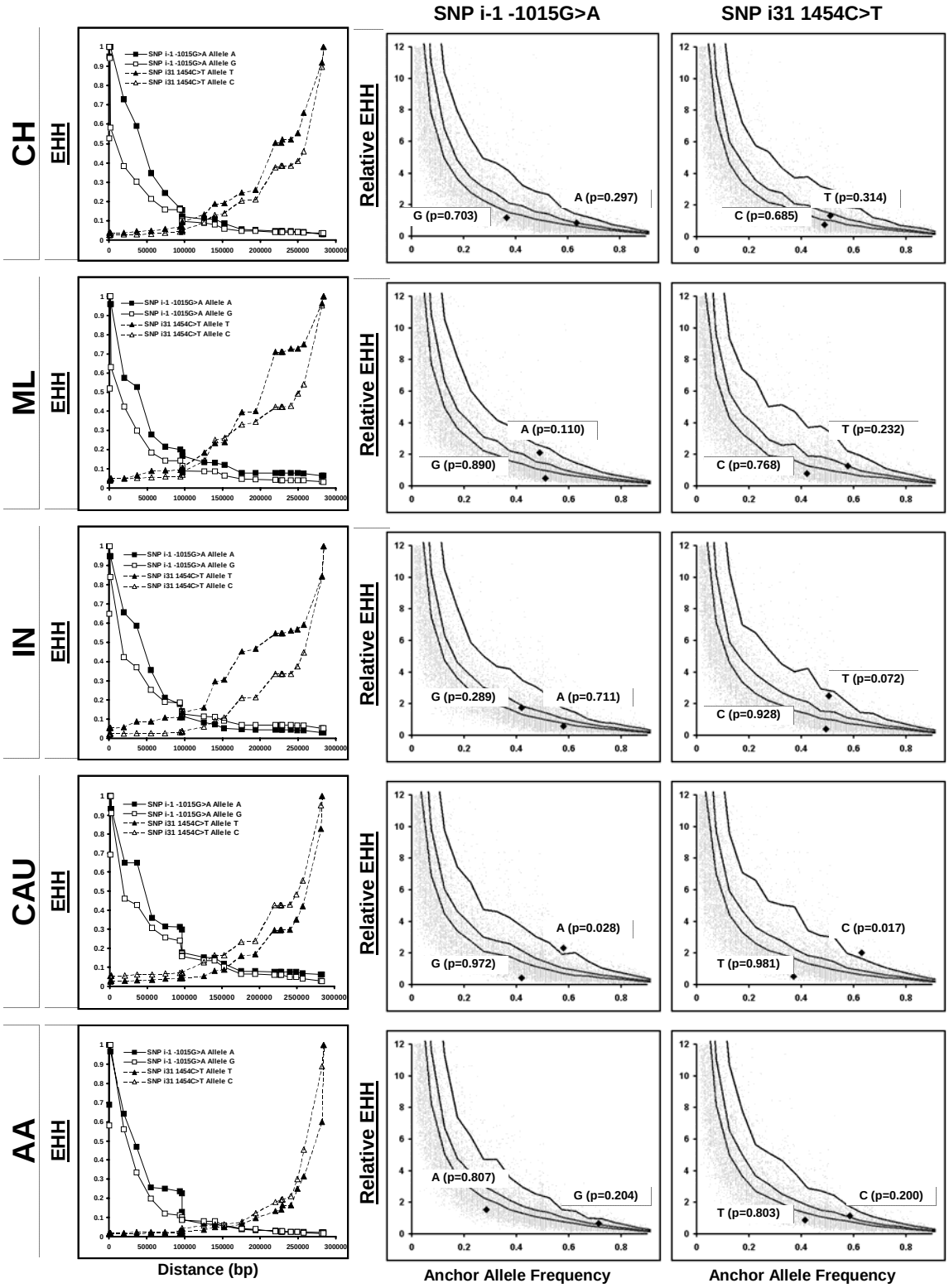


Figure 15. Extended haplotype homozygosity (EHH) and relative EHH (rEHH) plots of SNPs i-1 -1015G>A and i31 1454C>T. First vertical panel: EHH values at varying distances from SNPs i-1 -1015G>A and i31 1454C>T are plotted for all 5 populations. Second and third vertical panel: Relative EHH values at a most distant locus were plotted against allele frequency for each alternative allele of SNPs i-1 -1015G>A and i31 1454C>T respectively (represented as a black diamond), and compared against

simulated data (gray dots) under the structured population model with a recombination rate of 1.3 cM Mb^{-1}. The curved lines in each rEHH plot signify the 95th, 75th and 50th percentiles of the distribution of simulated alleles. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American).

The relative EHH value of SNP i-1 -1015 allele A of the European American population was 2.31 (Figure 15). Under the long-range haplotype test, this allele also demonstrated significant departure from evolutionary neutrality under four different population model assumptions and three different recombination rate assumptions ($P<0.05$) (Table 12). At another proximal SNP i-1 -527C>G which also displayed high LD scores with i-1 -1015G>A, the allele C showed a relative EHH value of 2.61. This allele in the European American population also showed significant evidence of positive selection under all tested population models and recombination rate assumptions except the structure model with the recombination rate 2.6 cM Mb^{-1} ($p = 0.06$) (Table 12). The C allele of the SNP in the 3' UTR region, i31 1454C>T had a high relative EHH value of 2.00. In the European American population, evidence of positive selection was also found for this allele under all tested models and assumptions. There were other SNP alleles showing high values of rEHH, but they failed to depart from all tested population models and recombination rate assumptions (data not shown). The Bonferroni adjustment was not applied to the test critical value despite using 25 polymorphic loci as the correction was too conservative for the LRH test and using an independent set of results, significant conclusions at the 5' flanking region of *ABCC*4 gene locus could still be replicated (as presented below).

i-1 -1015G>A			Constant Size			Expansion			Bottleneck			Structure	
Recombination rate		2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65
Population	Allele	1	2	3	6	7	8	9	10	11	12	13	14
CH	A	0.139576	**0.045897**	0.167024	0.070342	**0.004304**	0.173171	0.197002	0.062500	0.224784	0.280672	0.296761	0.251743
	G	0.860424	0.954103	0.832976	0.929658	0.995696	0.826829	0.802998	0.937500	0.775216	0.719328	0.703239	0.748257
ML	A	**0.003185**	**0.001359**	0.050420	**0.001754**	**0.000000**	**0.017857**	**0.040268**	**0.000000**	0.070953	0.109867	0.114544	0.086803
	G	0.996815	0.998641	0.949580	0.998246	1.000000	0.982143	0.959732	1.000000	0.929047	0.890133	0.885456	0.913197
IN	A	0.839015	0.880882	0.784173	0.864249	0.942128	0.788957	0.784934	0.870528	0.701117	0.711022	0.707567	0.685174
	G	0.160985	0.119118	0.215827	0.135751	0.057872	0.211043	0.215066	0.129472	0.298883	0.288978	0.292433	0.314826
CAU	A	**0.000000**	**0.000000**	**0.008726**	**0.000000**	**0.000000**	**0.001715**	**0.001715**	**0.000000**	**0.010846**	**0.028432**	**0.040129**	**0.024046**
	G	1.000000	1.000000	0.991274	1.000000	1.000000	0.998285	0.998285	1.000000	0.989154	0.971562	0.959871	0.975954
AA	A	0.931275	0.970160	0.894147	0.954248	0.997074	0.851852	0.865517	0.965919	0.806557	0.807000	0.824162	0.812214
	G	1.078244	**0.031185**	0.117160	0.054054	**0.003514**	0.151079	0.142703	**0.036565**	0.201248	0.203640	0.180581	0.186502

i-1 -527C>G			Constant Size			Expansion			Bottleneck			Structure	
Recombination rate		2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65
Population	Allele												
CH	C	0.161074	0.053790	0.186047	0.070240	**0.015458**	0.172260	0.233522	0.083562	0.252525	0.302958	0.317373	0.260690
	G	0.838926	0.946210	0.813953	0.929760	0.984542	0.827740	0.766478	0.916438	0.747475	0.697042	0.682627	0.739310
ML	C	**0.002853**	**0.000000**	**0.033520**	**0.001587**	**0.000000**	**0.012346**	**0.037363**	**0.007530**	0.080899	0.102478	0.115773	0.079252
	G	0.997147	1.000000	0.966480	0.998413	1.000000	0.987654	0.962637	0.992470	0.919101	0.897522	0.884227	0.920748
IN	C	0.812227	0.844700	0.732570	0.838150	0.921986	0.760391	0.755733	0.813107	0.650448	0.693331	0.698429	0.689491
	G	1.187773	0.155300	0.267430	0.161850	0.078014	0.239609	0.244267	0.186893	0.349552	0.306669	0.301633	0.310445
CAU	C	**0.000000**	**0.000000**	**0.007716**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.010081**	**0.032944**	0.062096	**0.028473**
	G	1.000000	1.000000	0.992284	1.000000	1.000000	1.000000	1.000000	1.000000	0.989919	0.967056	0.937904	0.971527
AA	C	0.932735	0.970569	0.877934	0.948515	0.996691	0.847269	0.869752	0.969831	0.809892	0.793822	0.809612	0.792067
	G	0.074010	**0.031847**	0.130435	0.055028	**0.004459**	0.154930	0.142132	**0.033557**	0.197932	0.222703	0.194619	0.211089

i31 1454C>T			Constant Size			Expansion			Bottleneck			Structure	
Recombination rate		2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65
Population	Allele												
CH	T	0.116477	**0.043210**	0.149211	**0.049383**	**0.000000**	0.194393	0.181672	0.089912	0.270665	0.314190	0.338559	0.282804
	C	0.883523	0.956790	0.850789	0.950617	1.000000	0.805607	0.818328	0.910088	0.729335	0.685240	0.661441	0.717196
ML	T	0.051896	**0.019272**	0.095238	**0.033413**	**0.000000**	0.094763	0.053012	**0.019108**	0.117647	0.232019	0.274822	0.204060
	C	0.948104	0.980728	0.904762	0.966587	1.000000	0.905237	0.946988	0.980892	0.882353	0.768160	0.725178	0.795940
IN	T	**0.000000**	**0.000000**	**0.005047**	**0.000613**	**0.000000**	**0.005587**	**0.006931**	**0.000000**	**0.036928**	0.071792	0.106089	0.058241
	C	1.000000	1.000000	0.994953	0.999387	1.000000	0.994413	0.993069	1.000000	0.963072	0.928208	0.893911	0.941759
CAU	T	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.990291	0.980947	0.964665	0.985908
	C	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.010163**	**0.017208**	**0.035725**	**0.013905**
AA	T	0.950125	0.987265	0.923246	0.961106	1.000000	0.901549	0.917148	0.968931	0.813846	0.803079	0.817536	0.804168
	C	0.053670	**0.017929**	0.076605	**0.038305**	**0.000000**	0.100642	0.086556	0.033333	0.193595	0.199549	0.203447	0.197437

Mindful that one of the SNPs, SNP i13/90G>A in the European American population deviated from HWE without Bonferroni correction and that prediction of different sets of haplotypes might lead to a false conclusion of positive selection, haplotypes previously generated were compared against those inferred from the set of 25 polymorphic markers sans SNP i13 90G>A. A total of 104 haplotypes were obtained for the latter set. The 2 most common haplotypes of both sets were congruent. Evidence of positive selection using the LRH test was reapplied on this newly inferred set of haplotypes. The SNP i-1 -1015G>A was tested. The relative EHH value of SNP i-1 -1015 allele A of the European American population was recalculated as 2.61 and using the LRH test, this allele seemed to appear on a single haplotype which again demonstrated significant departure from evolutionary neutrality under four different population model assumptions and three different recombination rate assumptions ($P<0.05$). The LRH test was therefore robust in detecting the evidence of positive selection in these cases.

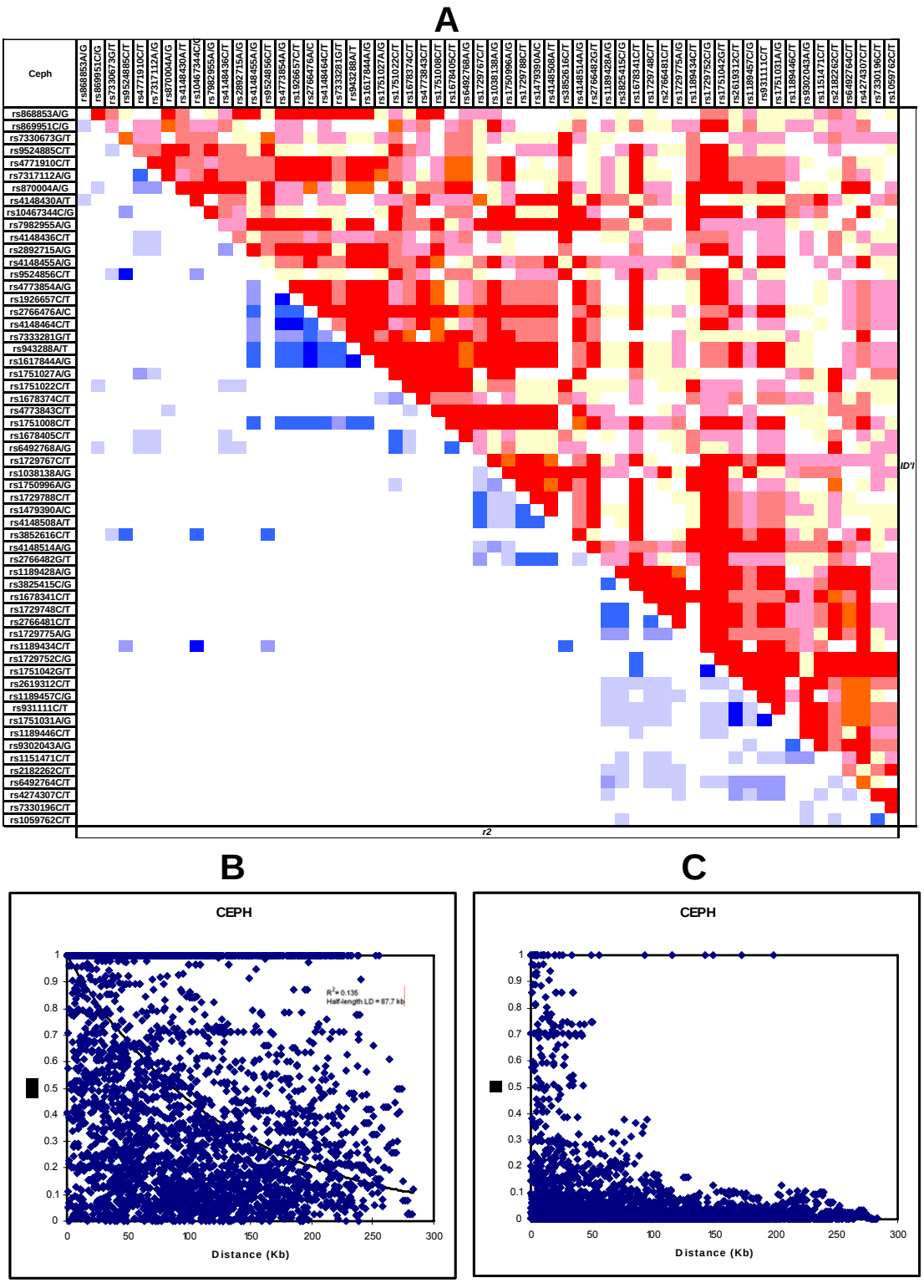


Figure 16. Pair-wise linkage disequilibrium profile of CEPH population in International Hapmap Project. Panel A. Pair-wise |*D'*| and r^2 values for highly polymorphic SNPs were summarized into tables. Panel B: |*D'*| between pairs of single SNPs was plotted against physical distance. Panel C: r^2 between pairs of single SNPs was calculated and plotted against physical distance.

The International Hapmap Project [http://www.hapmap.org/] is a large resource for studying verified SNPs in populations of different ethnicity. With genotype data from the CEPH (CEU), Han Chinese (HCB) and Yoruban (YRI) samples listed, comparative data could be obtained against those from this study (using populations of similar ancestry). Therefore CEPH population will be compared against the European American population, HCB against the Chinese population and finally YRI against the African Americans. Using the data set available publicly from the International HapMap Project, genotype data were obtained from 30 sets of CEPH trios (Utah residents with ancestry from northern and western Europe). Because of the similarity in ancestry between the CEPH trios and the European American population used in this study, analysis of the CEPH trios makes an interesting proposition. From the CEPH genotype data, the LD profile of 58 polymorphic SNPs surrounding the entire *ABCC*4 (from SNP i-1 -1393G>A (rs868853) to SNP i31 1454C>T (rs1059762)) showed that there were some differences from this study's own data (Figure 16). The half-length LD block and useful LD ($r^2 \geq 0.3$) were estimated to be 87.7 kb and 76.65 kb respectively. The discrepancy in LD measures between the 2 sets of data could be attributed to the large number of data points from the CEPH population, leading to a poorer correlation ($R^2=0.135$) in LD against genomic distance (as compared to a correlation factor of $R^2=0.208$ in the European American population from this study's data).

Due to the large number of the SNP markers listed in the International Hapmap Project, 29 polymorphic SNPs spanning approximately 140 kb from the 5' of *ABCC*4 were short-listed for haplotype analysis. Of these 29 SNPs, it was found that the G allele of the SNP rs869951 (SNP -1 -527C>G) possessed a high rEHH value of 2.74 at a frequency of 0.58. This SNP was subjected to the long range haplotype test and

demonstrated significant departure from evolutionary neutrality under all four different population model assumptions and three different recombination rate assumptions (P<0.05) (Figure 17A). This SNP lies in the 5' promoter region which is in high LD (r^2 = 0.92 see Figure 11) and is 488 bps away from i-1 -1015G>A (genotype data not available in CEPH data set), which was shown to pass the long range haplotype test in this study's data set. This simultaneous discovery of proximal SNPs showing positive selection in populations of similar ancestry is evident of genetic hitchhiking. When selection of one allele from either SNP occurs, the adjacent segment of DNA is passively selected with it, only later becoming separated by rare recombination. This gives rise to the tight linkage disequilibrium within the 5' region of *ABCC*4, which means that alleles may appear to be associated with variation in a phenotype without themselves causing that variation. As such this additional finding from the CEPH data gave greater confidence that the causal variant might be within the 5' region of *ABCC*4 gene locus.

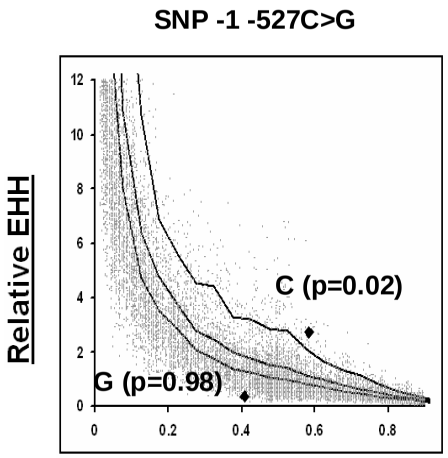
Similarly for the 3'end of the gene locus, 29 polymorphic SNPs from the International Hapmap Project spanning approximately 160 kb upstream of SNP i31 1454C>T (rs1059762) were analyzed. The major allele C of the SNP i31 1454C>T (rs1059762) was calculated to have a rEHH value of 1.00 at an allele frequency of 0.58. Here the p-value for a single structure model with recombination rate set at 1.3 cM Mb^{-1} is shown in Figure 17. Since no significant deviation was obtained for this model, no further analysis was performed. This discrepancy from the data used in this study was not surprising due to the fact that the HBD in the CEPH data showed a diffuse branching pattern at both alleles was too diffuse (Figure 17C).

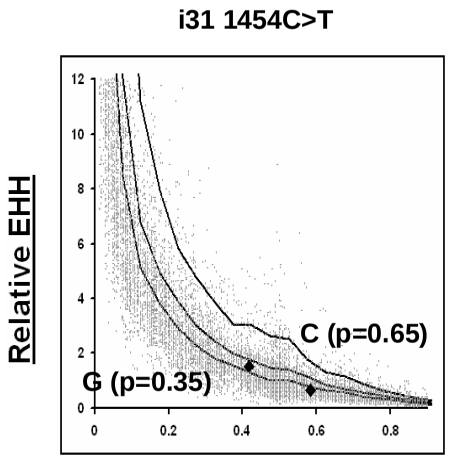
A

i-1 -527C>G rs869951		Constant Size			Expansion			Bottleneck			Structure		
Recombination rate		2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65
Population	Allele												
CEPH	C	0.000000	0.000000	0.002577	0.001826	0.000000	0.000887	0.005540	0.000000	0.003236	0.002736	0.014681	0.015532
	G	1.000000	1.000000	0.998309	0.998209	1.000000	0.999132	0.996390	1.000000	0.997868	0.997388	0.984671	0.991597

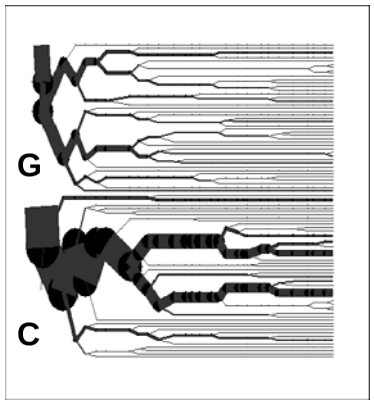
i31 1454C>T rs1059762		Constant Size			Expansion			Bottleneck			Structure		
Recombination rate		2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65	2.60	1.30	0.65
Population	Allele												
CEPH	C											0.647910	
	G											0.352090	

B





C



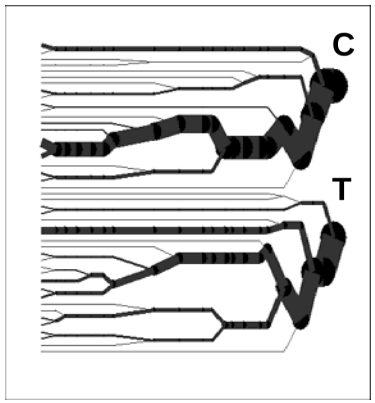


Figure 17. Long range haplotype test in CEPH data from International Hapmap Project. Panel A. P-values derived from comparing relative EHH values of 2 SNPs i-1 -527C>G and i31 1454C>T against simulated data points against 12 proposed population models. Panel B. relative EHH plots of SNP -1 -527C>G and SNP i31 1454C>T in CEPH population against allele frequency. Simulated data was generated from the proposed structure model with recombination rate set at 1.3 cM Mb^{-1}. Panel C. Haplotype Branching Diagrams at SNP -1 -527C>G (approximately 140 kb downstream) and at SNP i31 1454C>T (approximately 160 kb upstream)

Chapter 8 Genetic characterization of *ABCC*5 gene locus

8.1 Diversity of *ABCC*5 SNP allele frequencies amongst different populations

A ~1.2 kb fragment flanking the 5' end of *ABCC*5, which potentially contains the *ABCC*5 promoter, was sequenced from 30 individuals representing the three Asian ethnicities. Three novel SNPs (i-1 -889T>C; i-1 -177C>T and i-1 -115A>C) were identified and their frequencies determined in these populations (Table 13). However, each SNP was monomorphic in two of the three populations examined and present at very low frequency in the remaining population. These SNPs were excluded from further analysis.

Of 17 *ABCC*5 SNPs selected either from SNP databases or published reports that were genotyped in the 5 ethnic populations, 16 had high minor allele frequencies of >5% in all 5 populations (Table 13). Only intron 1 SNP 628G>C was monomorphic in European Americans and displayed low minor allele frequencies in Chinese (4.5%), Malays (2.9%), Indians and African Americans (both ~0.5%). The allele and genotype frequency distributions of all polymorphic SNPs were consistent with the Hardy-Weinberg equilibrium assumption (P>0.05) in at least four of the five populations examined (Table 13). When Bonferonni correction for multiple-locus testing was applied, no significant departure from Hardy-Weinberg equilibrium was detected for all genotyped SNPs in all populations.

The allele frequencies of most of the SNPs in the European American population that was genotyped were similar to frequencies of SNPs in another cohort of Caucasian population that was previously reported (Dazert et al. 2003) except for SNPs in exons 12 and 25. The differences in allele frequencies observed for SNPs in exons 12 and 25 could be due to stochastic variation as a result of the relatively small number of

samples genotyped (100 in this study and 20 in the previous study). It could also be due to differences in the population studied since the European American population this study examined was from the United States of America while those reported previously were probably German Caucasians (Dazert et al. 2003).

Allele frequencies at all SNP loci were similar between Chinese and Malays, and between European Americans and Indians ($P>0.05$). In sharp contrast, African American allele frequencies differed significantly from the non-African populations at most SNP loci ($P<0.05$; Table 13).

The significance of the variation in SNP allele frequencies between populations were evaluated using the F_{st} statistic (Akey et al. 2002). As shown in Table 13, the F_{st} values for all the 20 SNPs range from 0.02 to 0.14 with a mean value of 0.078. These values are all below the threshold of 0.45 suggesting that it is unlikely that any of the tested SNPs at this gene locus are under positive selection pressures.

SNP no.	SNP ID	dbSNP rs#	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's Exact P-value)					Pairwise Fst Value					Fst
													CH	ML	IN	CAU	AA	CH	ML	IN	CAU	AA	
								AA	GG	AG	A	G											
1	i-1-1990G>A		5' Flanking		CH	84	0.08	9.52	60.71	29.76	24.40	75.60		0.90	0.34	0.13	9.06E-08		0.00	0.00	0.01	0.14	0.06
					ML	86	0.35	8.14	58.14	33.72	25.00	75.00			0.40	0.17	2.95E-08			0.00	0.01	0.15	
					IN	91	0.72	7.69	49.45	42.86	29.12	70.88				0.58	1.01E-10				0.00	0.19	
					CAU	100	0.57	9.00	45.00	46.00	32.00	68.00					1.07E-12					0.21	
					AA	100	0.60	0.00	90.00	10.00	5.00	95.00											
								CC	TT	CT	C	T											
2	i-1-1821T>C		5' Flanking		CH	84	0.11	9.52	59.52	30.95	25.00	75.00		1.00	0.34	0.17	4.68E-08		0.00	0.00	0.01	0.15	0.06
					ML	86	0.35	8.14	58.14	33.72	25.00	75.00			0.34	0.17	2.95E-08			0.00	0.01	0.15	
					IN	91	0.61	7.69	48.35	43.96	29.67	70.33				0.66	4.88E-11				0.00	0.19	
					CAU	100	0.57	9.00	45.00	46.00	32.00	68.00					1.07E-12					0.21	
					AA	100	0.60	0.00	90.00	10.00	5.00	95.00											
								AA	TT	AT	A	T											
3	i-1-1679T>A		5' Flanking		CH	84	1.00	10.71	45.24	44.05	32.74	67.26		0.21	2.42E-05	6.65E-08	0.16		0.00	0.09	0.14	0.01	0.10
					ML	86	0.04	20.93	41.86	37.21	39.53	60.47			3.54E-08	2.68E-11	1.00			0.15	0.21	0.00	
					IN	91	0.13	0.00	72.53	27.47	13.74	86.26				0.27	6.46E-09				0.00	0.16	
					CAU	100	0.27	0.00	80.00	20.00	10.00	90.00					2.83E-12					0.21	
					AA	100	0.68	17.00	37.00	46.00	40.00	60.00											
								CC	TT	CT	C	T											
4	i-1-1205C>T		5' Flanking		CH	84	0.66	7.14	57.14	35.71	25.00	75.00		0.24	4.75E-04	1.35E-07	0.62		0.00	0.06	0.14	0.00	0.09
					ML	86	0.91	3.49	65.12	31.40	19.19	80.81			2.15E-06	5.69E-11	0.45			0.12	0.20	0.00	
					IN	91	0.04	13.19	27.47	59.34	42.86	57.14				0.08	2.81E-05				0.01	0.09	
					CAU	100	0.11	23.00	19.00	58.00	52.00	48.00					1.39E-09					0.17	
					AA	100	0.97	5.00	60.00	35.00	22.50	77.50											
								AA	CC	AC	A	C											
5	i-1-889T>C		5' Flanking		CH	84	0.87	10.71	44.05	45.24	33.33	66.67		0.26	1.45E-05	3.51E-08	9.95E-11		0.00	0.10	0.15	0.20	0.13
					ML	86	0.04	20.93	41.86	37.21	39.53	60.47			3.54E-08	2.68E-11	3.68E-14			0.15	0.21	0.26	
					IN	91	0.13	0.00	72.53	27.47	13.74	86.26				0.27	0.04				0.00	0.02	
						100	0.27	0.00	80.00	20.00	10.00	90.00					0.37					0.00	
						100	0.45	0.00	86.00	14.00	7.00	93.00											
								CC	TT	CT	C	T											
6	i-1-793C>A		5' Flanking		CH	84	0.84	77.38	1.19	21.43	88.10	11.90		0.35	1.56E-09	5.43E-12	9.98E-05		0.00	0.18	0.22	0.08	0.09
					ML	86	0.08	68.60	0.00	31.40	84.30	15.70			3.05E-07	3.09E-09	4.08E-03			0.13	0.17	0.04	
					IN	91	0.11	31.87	12.09	56.04	59.89	40.11				0.47	0.02				0.00	0.02	
					CAU	100	0.17	28.00	16.00	56.00	56.00	44.00					1.76E-03					0.05	
					AA	100	0.36	53.00	10.00	37.00	71.50	28.50											
								CC	TT	CT	C	T											
7	i-1-177C>T		5' Flanking		CH	84	0.75	47.62	10.71	41.67	68.45	31.55		0.21	6.49E-05	1.18E-06	2.13E-05		0.00	0.08	0.12	0.09	0.09
					ML	86	0.13	41.86	18.60	39.53	61.63	38.37			1.17E-07	4.86E-10	1.78E-08			0.14	0.18	0.15	
					IN	91	0.13	72.53	0.00	27.47	86.26	13.74				0.44	0.88				0.00	0.00	
						100	0.22	78.00	0.00	22.00	89.00	11.00					0.64					0.00	
						100	0.25	77.00	3.00	20.00	87.00	13.00											

Table 13. Allele frequencies of the different SNPs in the different ethnic populations

SNP no.	SNP ID	dbSNP rs#	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's Exact P-value)				Pairwise Fst Value					Fst
													CH	ML	IN	CAU	AA	CH	ML	IN	CAU	AA	
								CC	TT	CT	C	T											
8	i-1 -115A>C		5' Flanking		CH	84	0.58	32.14	21.43	46.43	55.36	44.64		0.44	0.59	0.17	2.65E-04		0.00	0.00	0.01	0.07	0.02
					ML	86	0.33	38.37	18.60	43.02	59.88	40.12			0.83	0.67	4.01E-03			0.00	0.00	0.04	
					IN	91	0.22	30.77	14.29	54.95	58.24	41.76				0.40	1.60E-03				0.00	0.05	
						100	0.69	40.00	15.00	45.00	62.50	37.50					0.02					0.03	
						100	0.69	54.00	6.00	40.00	74.00	26.00											
								AA	GG	AG	A	G											
9	i1 628G>C	rs4148556	Intron 1		CH	84	0.55	30.95	22.62	46.43	54.17	45.83		0.44	0.45	0.11	4.86E-05		0.00	0.00	0.01	0.08	0.02
					ML	86	0.55	36.05	18.60	45.35	58.72	41.28			1.00	0.46	1.34E-03			0.00	0.00	0.05	
					IN	91	0.22	30.77	14.29	54.95	58.24	41.76				0.40	1.07E-03				0.00	0.05	
					CAU	100	0.69	40.00	15.00	45.00	62.50	37.50					0.01					0.03	
					AA	100	0.79	55.00	6.00	39.00	74.50	25.50											
								CC	TT	CT	C	T											
10	i1 1834C>T	rs4148557	Intron 1		CH	84	0.76	76.19	1.19	22.62	87.50	12.50		0.64	4.60E-08	5.75E-11	4.43E-04		0.00	0.15	0.20	0.06	0.08
					ML	86	0.11	70.93	0.00	29.07	85.47	14.53			6.14E-07	1.22E-09	2.42E-03			0.13	0.17	0.04	
					IN	91	0.17	35.16	10.99	53.85	62.09	37.91				0.35	0.04				0.00	0.02	
					CAU	100	0.31	30.00	16.00	54.00	57.00	43.00					1.64E-03					0.05	
					AA	100	0.78	52.00	7.00	41.00	72.50	27.50											
								AA	GG	AG	A	G											
11	i2 7980C>T	rs2292997	Intron 2		CH	84	0.22	76.19	0.00	23.81	88.10	11.90		0.63	2.05E-07	1.21E-12	4.36E-03		0.00	0.14	0.23	0.04	0.10
					ML	86	0.13	72.09	0.00	27.91	86.05	13.95			2.51E-06	3.18E-11	0.02			0.11	0.20	0.02	
					IN	91	0.04	36.26	7.69	56.04	64.29	35.71				0.06	9.75E-03				0.01	0.03	
					CAU	100	0.06	25.00	16.00	59.00	54.50	45.50					5.37E-06					0.10	
					AA	100	0.77	58.00	5.00	37.00	76.50	23.50											
								AA	GG	AG	A	G											
12	i5 374C>T	rs3749438	Intron 5		CH	84	0.76	42.86	13.10	44.05	64.88	35.12		0.43	2.98E-06	1.26E-08	3.53E-13		0.00	0.11	0.16	0.25	0.14
					ML	86	0.11	40.70	19.77	39.53	60.47	39.53			3.54E-08	4.40E-11	3.72E-15			0.15	0.20	0.29	
					IN	91	0.13	72.53	0.00	27.47	86.26	13.74				0.35	7.91E-03				0.00	0.03	
					CAU	100	0.24	79.00	0.00	21.00	89.50	10.50					0.10					0.01	
					AA	100	0.56	89.00	0.00	11.00	94.50	5.50											
								CC	TT	CT	C	T											
13	e8 1145A>G		Exon 8	Gln382Gln	CH	84	0.05	29.76	30.95	39.29	49.40	50.60		0.45	1.34E-10	1.61E-10	2.03E-08		0.00	0.21	0.20	0.15	0.11
					ML	86	0.42	31.40	23.26	45.35	54.07	45.93			1.71E-08	2.98E-08	1.95E-06			0.16	0.15	0.11	
					IN	91	1.00	67.03	3.30	29.67	81.87	18.13				0.90	0.31				0.00	0.00	
					CAU	100	0.80	66.00	4.00	30.00	81.00	19.00					0.46					0.00	
					AA	100	0.54	59.00	4.00	37.00	77.50	22.50											
								CC	TT	CT	C	T											
14	e9 1185C>T	rs1132776	Exon 9	Ala395Ala	CH	84	0.57	29.76	17.86	52.38	55.95	44.05		0.83	2.94E-07	2.45E-07	7.02E-06		0.00	0.14	0.13	0.10	0.08
					ML	86	0.57	31.40	22.09	46.51	54.65	45.35			6.54E-08	5.39E-08	1.83E-06			0.15	0.15	0.11	
					IN	91	0.57	67.03	4.40	28.57	81.32	18.68				1.00	0.45				0.00	0.00	
					CAU	100	0.80	66.00	4.00	30.00	81.00	19.00					0.54					0.00	
					AA	100	0.28	59.00	3.00	38.00	78.00	22.00											

Table 13 (continued).

SNP no.	SNP ID	dbSNP rs#	Location	Amino Acid Change	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences (Fisher's Exact P-value)					Pairwise Fst Value					Fst
													CH	ML	IN	CAU	AA	CH	ML	IN	CAU	AA	
								AA	CC	AC	A	C											
15	e12 1782C>T	rs939336	Exon 12	Cys594Cys	CH	84	0.58	60.71	3.57	35.71	78.57	21.43		0.21	0.07	1.72E-05	2.71E-08		0.00	0.01	0.09	0.15	0.05
					ML	86	0.08	68.60	0.00	31.40	84.30	15.70			0.65	3.46E-03	1.53E-05			0.00	0.04	0.09	
					IN	91	0.13	72.53	0.00	27.47	86.26	13.74				0.01	1.21E-04				0.03	0.07	
					CAU	100	0.52	88.00	0.00	12.00	94.00	6.00					0.23					0.01	
					AA	100	0.76	94.00	0.00	6.00	97.00	3.00											
								CC	TT	CT	C	T											
16	e25 3624C>T	rs3749442	Exon 25	Leu1208Leu	CH	84	0.74	29.76	19.05	51.19	55.36	44.64		0.91	6.53E-07	1.59E-08	2.36E-06		0.00	0.13	0.16	0.11	0.09
					ML	86	0.57	31.40	22.09	46.51	54.65	45.35			2.66E-07	5.92E-09	1.50E-06			0.13	0.16	0.12	
					IN	91	0.77	64.84	4.40	30.77	80.22	19.78				0.60	0.71				0.00	0.00	
					CAU	100	0.97	68.00	3.00	29.00	82.50	17.50					0.38					0.00	
					AA	100	0.34	60.00	3.00	37.00	78.50	21.50											
								AA	GG	AG	A	G											
17	e30 4896G>A	rs3749445	Exon 30		CH	84	0.05	19.05	20.24	60.71	49.40	50.60		0.13	0.91	0.83	5.19E-08		0.01	0.00	0.00	0.15	0.07
					ML	86	0.08	38.37	22.09	39.53	58.14	41.86			0.07	0.06	9.43E-13			0.01	0.02	0.24	
					IN	91	0.34	20.88	24.18	54.95	48.35	51.65				1.00	6.36E-08				0.00	0.14	
					CAU	100	0.41	21.00	25.00	54.00	48.00	52.00					7.00E-08					0.13	
					AA	100	0.21	7.00	63.00	30.00	22.00	78.00											
								AA	GG	AG	A	G											
18	e30 5557A>G	rs562	Exon 30		CH	84	0.05	20.24	19.05	60.71	50.60	49.40		0.04	0.91	1.00	2.10E-03		0.02	0.00	0.00	0.05	0.02
					ML	86	0.07	19.77	41.86	38.37	38.95	61.05			0.02	0.03	0.39			0.03	0.02	0.00	
					IN	91	0.34	24.18	20.88	54.95	51.65	48.35				0.84	9.00E-04				0.00	0.05	
					CAU	100	0.16	22.00	21.00	57.00	50.50	49.50					1.68E-03					0.05	
					AA	100	0.03	7.00	38.00	55.00	34.50	65.50											
								AA	GG	AG	A	G											
19	i30 6272A>G	rs1000002	3' Flanking		CH	84	0.08	21.43	19.05	59.52	51.19	48.81		0.23	0.92	0.40	4.21E-08		0.00	0.00	0.00	0.15	0.07
					ML	86	8.60E-03	40.70	24.42	34.88	58.14	41.86			0.17	0.03	1.17E-11			0.01	0.02	0.22	
					IN	91	0.60	24.18	23.08	52.75	50.55	49.45				0.47	4.60E-08				0.00	0.14	
					CAU	100	0.15	18.00	25.00	57.00	46.50	53.50					2.06E-06					0.11	
					AA	100	0.79	6.00	59.00	35.00	23.50	76.50											
								AA	GG	AG	A	G											
20	i30 7161G>A	rs1533682	3' Flanking		CH	84	0.59	2.38	76.19	21.43	13.10	86.90		1.00	1.90E-07	1.55E-10	0.05		0.00	0.14	0.19	0.02	0.09
					ML	86	0.15	0.00	73.26	26.74	13.37	86.63			2.28E-07	1.92E-10	0.06			0.14	0.19	0.01	
					IN	91	0.10	9.89	35.16	54.95	37.36	62.64				0.30	4.55E-04				0.00	0.06	
					CAU	100	0.31	16.00	30.00	54.00	43.00	57.00					3.46E-06					0.10	
					AA	100	0.72	5.00	63.00	32.00	21.00	79.00											

Table 13 (continued).

8.2 Strong LD at the *ABCC5* gene locus in all five populations

Linkage disequilibrium across the 100 kb *ABCC5* genomic region was examined in the five ethnic populations for all possible SNP pairs. Only the 16 SNPs, with minor allele frequency of >5% in all populations were included in this analysis. The two common LD statistics, $|D'|$ and r^2, which measure recombination rate and associative power, respectively, were determined (Shifman et al. 2003).

All five populations showed generally strong LD across the entire gene locus with high $|D'|$ and r^2 values between the fourth (intron -1 -1205C>T) and last (intron 30 7161G>A) SNPs ($p<0.05$) (Figure 18). Similar to observations at other gene loci (Johnson et al. 2001; Tang et al. 2004), LD at the *ABCC5* gene locus generally decreased with physical distance although the correlation was weak ($R^2<0.32$) (Figure 18B). Half-length LD ($LD_{0.5}$) was shortest in African Americans (106 kb), followed by the Malays (144 kb), Chinese (165 kb), European Americans (266 kb), and Indians (293 kb). When $r^2 \geqslant 0.3$ was utilized as a threshold for useful LD in association studies (Pritchard and Przeworski 2001), useful LD was found to extend beyond the entire 100 kb *ABCC5* gene locus in all five populations (Figure 18C).

As illustrated in Figure 19, less intensive branching is seen in the 4 non-African populations compared to the African American population, indicating that LD decay across the *ABCC5* gene is more pronounced in the African American population. Interestingly, the Chinese and Malay *ABCC5* genes share very similar branching patterns, while the European Americans and Indians share a similar pattern. These results indicate that LD decay is similar between Chinese and Malays and between European Americans and Indians.

A
B
C
I
II
III
CH
ML
IN
r2
D'
R²=0.317
Half-length LD=165.0 kb
R²=0.319
Half-length LD=144.2 kb
R²=0.101
Half-length LD=293.3 kb
Distance

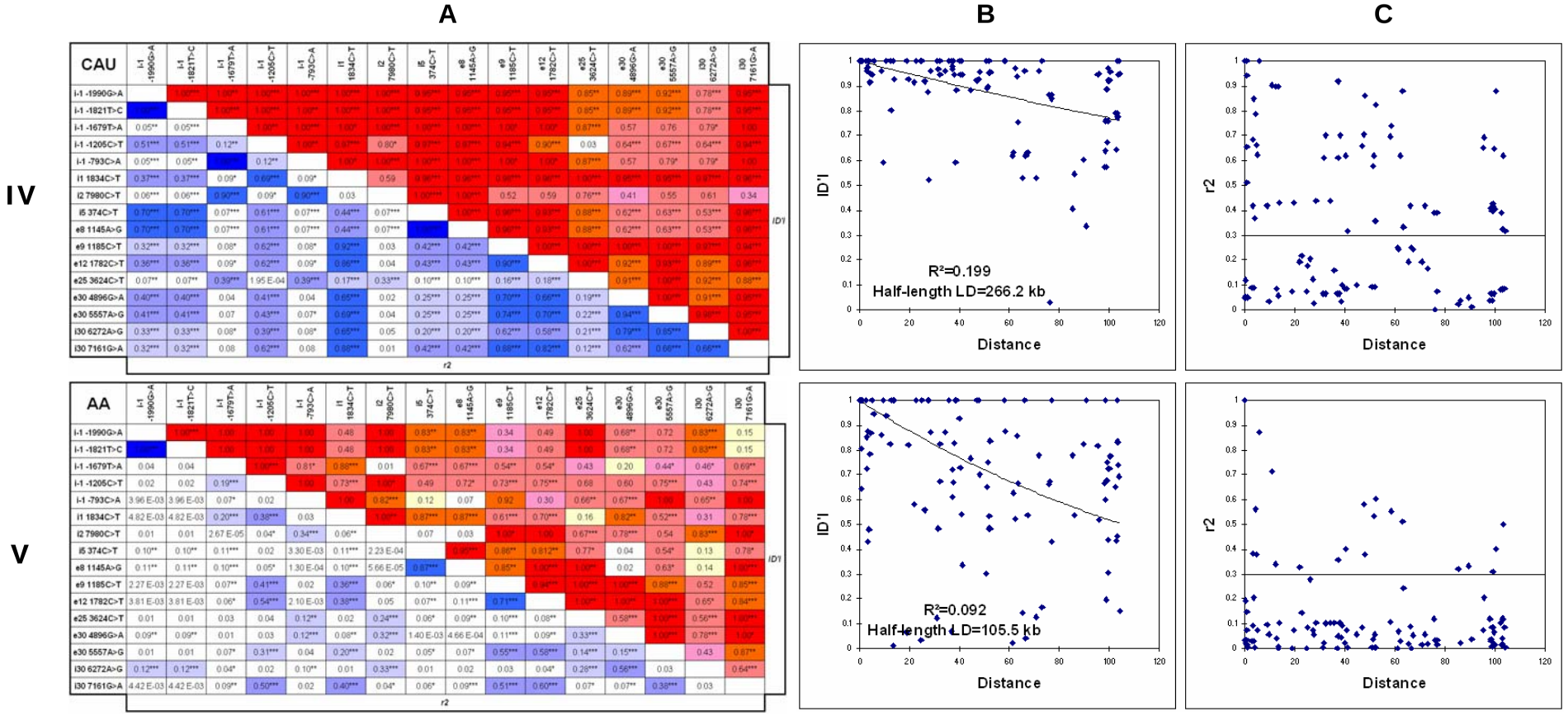


Figure 18. Pair-wise linkage disequilibrium profiles of *ABCC5* gene locus.
Panel A: $|D'|$, r^2 as well as Fisher's exact p value of association were calculated for each pair of the 16 highly polymorphic SNPs (>5%) in five populations and summarized into tables [I-Chinese (CH); II-Malay (ML); III-Indian (IN); IV-European American (CAU); V-African American (AA)]. Cells in the triangular section above the diagonal indicate $|D'|$ values, whereas those in the lower triangular section contain the r^2 results. Gradient of either red ($|D'|$) or blue (r^2) color for each cell is presented according to the scale specified beside the tables and graphs, with darker colours representing higher value of $|D'|$ 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, 0.001 and 0.01, and below 0.001specified by a single-asterisk, double-asterisk and triple-asterisk, respectively. "-" indicates pairs where calculations of LD cannot be applied. Panel B: $|D'|$ between pairs of single SNPs was calculated (see Methods) and plotted against physical distance in the five populations. The correlation coefficient R^2, as well as the half-length LD distances are shown next to the fitted line. Panel C: r^2 between pairs of single SNPs was calculated (see Methods) and plotted against physical distance in the five populations. The threshold for useful LD is set at $r2 \geqslant 0.3$.

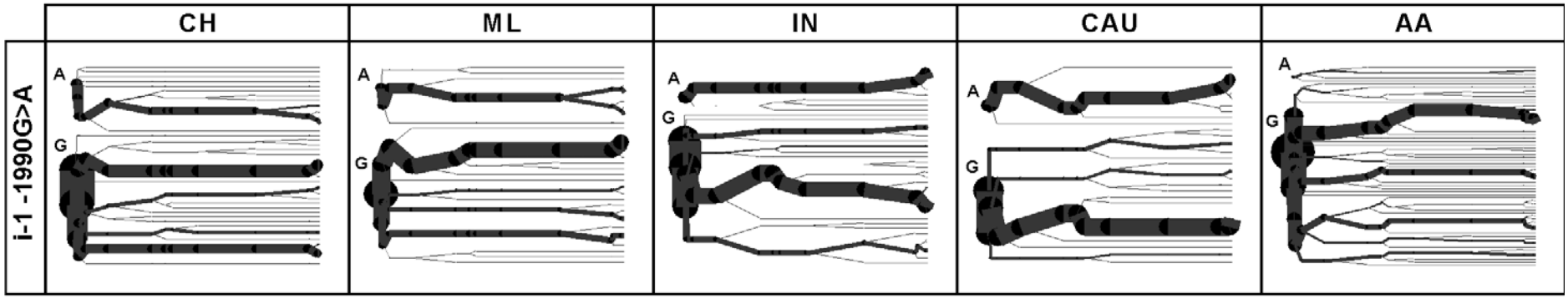


Figure 19. LD decay profile at the *ABCC5* locus.
SNP intron -1 -1990G>A, representing the 5' end of the *ABCC5* gene locus was selected as the root or core marker locus and denoted by a black dot. The position of each locus along the X-axis is scaled to the physical distance between each locus. Each SNP locus downstream of the root locus is denoted as a node specified by a black dot. Grey lines are drawn between nodes to represent the specific haplotypes. As the LD decays progressively with increasing distance downstream of the test locus, branching of the diagram depends on the presence of alternative alleles at the next consecutive node. The thickness of the branches corresponds to the frequency of the haplotype while the size of the black dots corresponds to the allele frequency. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American).

8.3 Low but varied *ABCC*5 haplotype diversity among the five populations

Assuming maximum random association between the 16 polymorphic *ABCC*5 SNPs, the number of haplotypes predicted based on the tested sample size and observed allele frequencies in a simulated population iterated 1,000 times was found to be 164.57 ± 1.63, 168.93 ± 1.68, 179.98 ± 1.29, 197.41 ± 1.61 and 177.77 ± 3.53 in the Chinese, Malay, Indian, European American, and African American populations, respectively. If no recombination occurs, the number of possible haplotypes would be decreased to 17. In this study, 35, 28, 32, 26 and 55 haplotypes were observed in the Chinese, Malay, Indian, European American, and African American populations, respectively. Hence, the observed number of haplotypes in the Chinese, Malay, Indian, European American, and African American groups constituted only 21.27%, 16.57%, 17.78%, 13.17%, and 30.94% of expected haplotypes, simulated under maximum random association, respectively. These results suggest that only limited recombination has occurred between SNP pairs at this locus. The greatest haplotype diversity was observed in the African American population (Figure 20).

In total, 130 different *ABCC*5 haplotypes were observed from these five populations. Several of the haplotypes are specific to the Chinese (22 haplotypes), Malay (10), Indian (17), European American (16) and African American (40) populations.

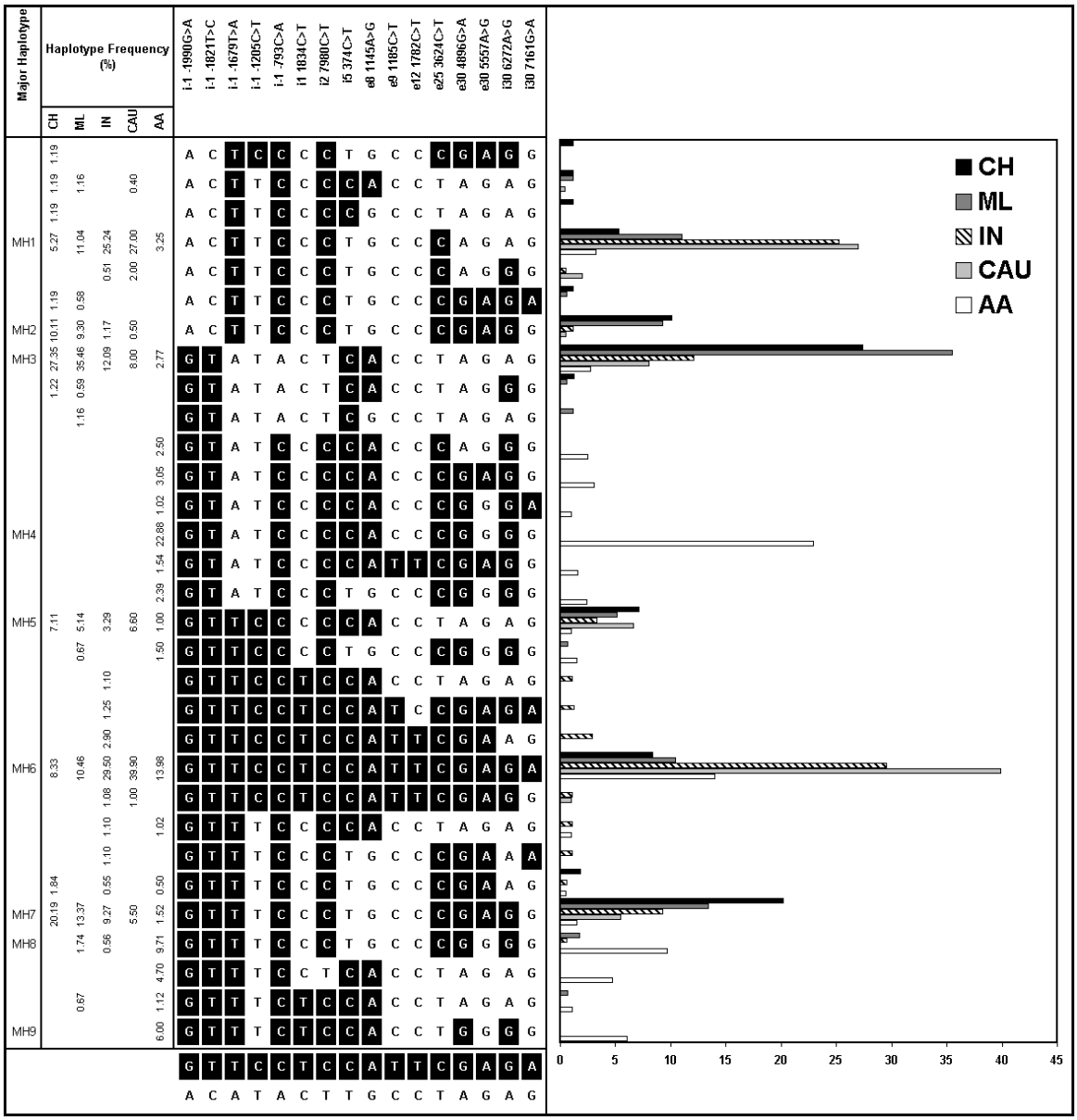


Figure 20. Haplotype profiles of SNPs at the *ABCC5* gene locus. Expectation-Maximization algorithm was utilized to infer haplotype frequencies from genotype data of the 16 highly polymorphic SNPs at the *ABCC5* gene locus. A total of 130 haplotypes occur in at least one population, of which 35, 28, 32, 26 and 55 were found in the Chinese, Malay, Indian, European American and African American groups, respectively. Only haplotypes occurring at >1% are presented in this figure and those with frequencies >5% are labeled as MH1-MH9. Haplotype profiles are represented as horizontal arrays of black or white boxes with the allele specified. Each column of boxes represents a SNP locus. Each row of boxes represents a haplotype, with the estimated frequencies expressed as a percentage shown in the adjacent table as well as portrayed in a bar graph. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American).

Figure 20 shows the haplotypes that occur at higher than 1% frequency in at least one population. While the most common haplotype accounts for 27-40% of total alleles in each 4 non-African population, the most common haplotype in the African American population accounts for only 22.9%. Of the five major haplotypes that occur in all five populations, four show similarities in frequency distribution between the Chinese and Malay and between the Indians and European Americans. Fisher's Exact Test showed no significant difference in haplotype frequencies between Indians and European Americans ($p=0.14456$), as well as between Chinese and Malays ($p=0.06915$), but highly significant difference for all other pair-wise comparisons between populations ($p<0.00001$).

8.4 Tagging SNPs at the *ABCC*5 gene locus

The above results suggest that the entire 100 kb of the *ABCC*5 gene locus may reside within a region of strong LD. To reduce genotyping costs and effort in future association studies, it would be useful to identify tSNPs that represent the majority of observed haplotypes but with minimal reduction in power to detect associations. The "Tag-IT" program (Weale et al. 2003) was utilized to identify tSNPs at the *ABCC*5 gene locus. In the non-African populations, a minimum of 4 tSNPs were sufficient to generate haplotype r^2 values of ~0.9 and thus account for ~90% of the observed haplotypes in each population (Table 14). However, the panel of 4 tSNP combinations with haplotype r^2 values of ~0.9 differed among the four populations, suggesting that the panel of tSNPs identified in one population may not be suitable for use in other populations without a reduction of power (Table 14). Nonetheless, more than 40 different combinations of 4 tSNPs with haplotype r^2 values >0.9 could be utilized for association studies in any of the non-African populations. In the African American population, at least six tSNPs are required to represent 90% of all the observed haplotypes (Table 14).

Number of tSNPs	i-1 -1990G>A	i-1 -1821T>C	i-1 -1679T>A	i-1 -1205C>T	i-1 -793C>A	i1 1834C>T	i2 7980C>T	i5 374C>T	e8 1145A>G	e9 1185C>T	e12 1782C>T	e25 3624C>T	e30 4896G>A	e30 5557A>G	i30 6272A>G	i30 7161G>A	Possible tSNP sets	CH	ML	IN	CAU	AA
4	■		■							■			■				1820	0.9140	0.9378	0.9094	0.9384	
	■		■							■				■				0.9140	0.9165	0.9029	0.9432	
	■		■	■									■					0.9129	0.9131	0.9067	0.9369	
	■		■	■										■				0.9129	0.9187	0.9125	0.9461	
	■		■					■					■					0.9216	0.9482	0.9148	0.9529	
	■		■					■						■				0.9216	0.9546	0.9237	0.9606	
	■		■						■					■				*0.8961*	**0.9591**	0.9277	0.9606	
	■			■								■	■					0.9263	0.9052	0.9016	0.9474	
	■			■								■		■				0.9263	0.9174	0.9041	0.9551	
	■			■								■			■			0.9300	0.9010	0.9089	0.9476	
	■			■									■			■		0.9010	0.9147	0.9112	0.9415	
	■			■										■	■			0.9203	0.9062	0.9191	0.9544	
	■			■	■								■					0.9082	0.9131	0.9067	0.9369	
	■			■										■				0.9082	0.9187	0.9125	0.9461	
	■			■			■							■				0.9070	0.9044	0.9125	0.9491	
	■			■				■					■					**0.9305**	0.9352	0.9062	0.9529	
	■			■				■						■				**0.9305**	0.9355	0.9130	0.9606	
	■			■					■				■					0.9214	0.9438	0.9046	0.9529	
	■			■					■					■				0.9214	0.9388	0.9114	0.9606	
	■			■							■				■			*0.8414*	0.9236	*0.8790*	**0.9627**	
	■				■					■			■					0.9222	0.9378	0.9094	0.9384	
	■				■					■				■				0.9222	0.9165	0.9029	0.9432	
	■				■	■							■					0.9030	0.9273	0.9145	0.9529	
	■				■	■								■				0.9030	0.9140	0.9151	0.9585	
	■				■			■					■					0.9218	0.9482	0.9148	0.9529	
	■				■			■						■				0.9218	0.9546	0.9237	0.9606	
	■				■			■							■			0.9046	0.9171	0.9073	0.9514	
	■				■				■					■				*0.8945*	**0.9591**	0.9277	0.9606	
		■	■						■					■				*0.8956*	**0.9591**	0.9300	0.9606	
		■	■							■			■					0.9100	0.9378	0.9054	0.9384	
		■	■	■									■					0.9119	0.9131	0.9027	0.9369	
		■	■	■										■				0.9119	0.9187	0.9090	0.9461	
		■	■					■					■					0.9204	0.9482	0.9141	0.9529	
		■	■					■						■				0.9204	0.9546	0.9227	0.9606	
		■		■								■	■					0.9227	0.9052	0.9068	0.9474	
		■		■								■		■				0.9227	0.9174	0.9093	0.9551	
		■		■								■			■			0.9289	0.9010	0.9145	0.9476	
		■		■										■	■			0.9193	0.9062	0.9180	0.9544	
		■		■	■								■					0.9100	0.9131	0.9027	0.9369	
		■		■	■									■				0.9100	0.9187	0.9090	0.9461	
		■		■			■							■				0.9088	0.9044	0.9090	0.9491	
		■		■				■					■					0.9270	0.9352	0.9081	0.9529	
		■		■				■						■				0.9270	0.9355	0.9150	0.9606	
		■		■					■					■				0.9179	0.9388	0.9166	0.9606	
		■		■							■				■			*0.8339*	0.9236	*0.8790*	**0.9627**	
		■			■				■					■				*0.8066*	**0.9591**	0.9300	0.9606	
		■			■					■			■					0.9176	0.9378	0.9054	0.9384	
		■			■			■					■					0.9202	0.9482	0.9141	0.9529	
		■			■			■						■				0.9202	0.9546	0.9227	0.9606	
		■			■			■							■			0.9031	0.9171	0.9079	0.9514	
		■			■			■								■		0.9031	0.9171	0.9079	0.9514	
			■						■				■			■		0.8255	0.9034	**0.9430**	0.9603	
					■				■				■			■		0.8241	0.9034	**0.9430**	0.9603	
							■		■				■			■		0.7998	0.8921	**0.9403**	0.9562	
6			■			■		■		■				■	■							**0.9111**
			■			■		■					■	■	■							0.9109
			■	■					■	■				■	■							0.9087
			■			■			■	■				■	■							0.9081
			■	■				■					■	■	■							0.9080
			■			■			■				■	■	■							0.9077
			■	■		■		■						■	■							0.9071
			■	■				■		■				■	■							0.9071
			■					■			■		■	■	■		8008					0.9064
			■	■					■				■	■	■							0.9049
			■					■		■					■	■						0.9046
			■					■					■	■	■	■						0.9031
			■			■		■						■	■	■						0.9027
			■						■	■				■	■	■						0.9020
			■	■		■			■					■	■							0.9016
			■			■	■		■					■	■							0.9005
			■					■			■		■		■							0.9002

Bold - highest haplotype r^2 value for the number of tSNPs in each population.
Italics - panel of 4 tSNPs that has haplotype r^2 value of <0.9 in each of the non-African populations.

Table 14. Tagging SNPs at the *ABCC5* gene locus in 5 ethnic populations.

8.5 Evidence of positive selection not apparent at the *ABCC*5 gene locus

Utilizing a modified long-range haplotype method proposed by Sabeti et al, the relative EHH were calculated for their respective allele frequencies to test for evidence of recent positive selection at *ABCC*5 gene locus (Table 15). The observation that *ABCC*5 SNPs with high rEHH and low allele frequencies (as compared to *ABCC*4 SNPs) runs contrary to the premise of recent positive selection. We were also unable to obtain enough simulated data for comparison as the entire *ABCC*5 gene locus is in a region of strong linkage disequilibrium and the genomic distance covering all the SNPs in *ABCC*5 gene locus was too small. Only the first and last SNPs could be effectively tested.

SNP	CH		ML		IN		CAU		AA	
	Allele Frequency	rEHH	Allele Frequency	rEHH	Allele Frequency	rEHH	Allele Frequency	rEHH	Allele Frequency	rEHH
i-1 -1990G>A										
A	0.244	1.210	0.250	1.340	0.291	**3.304**	0.320	1.833	0.050	**4.244**
G	0.756	0.826	0.750	0.746	0.709	0.303	0.680	0.546	0.950	0.236
i-1 -1821T>C										
C	0.250	1.138	0.250	1.340	0.297	**3.136**	0.320	1.833	0.050	**4.244**
T	0.750	0.879	0.750	0.746	0.703	0.319	0.680	0.546	0.950	0.236
i-1 -1679T>A										
T	0.673	0.222	0.605	0.189	0.863	0.297	0.900	0.472	0.600	0.276
A	0.327	**4.504**	0.395	**5.305**	0.137	**3.372**	0.100	**2.118**	0.400	**3.617**
i-1 -1205C>T										
C	0.250	0.918	0.192	1.457	0.571	0.532	0.480	0.568	0.775	0.343
T	0.750	1.089	0.808	0.686	0.429	1.880	0.520	1.761	0.225	**2.917**
i-1 -793C>A										
C	0.667	0.243	0.605	0.189	0.863	0.300	0.900	0.472	0.930	0.617
A	0.333	**4.115**	0.395	**5.283**	0.137	**3.332**	0.100	**2.118**	0.070	1.621
i1 1834C>T										
C	0.881	0.372	0.843	0.475	0.599	0.444	0.560	0.323	0.715	0.580
T	0.119	**2.692**	0.157	**2.105**	0.401	**2.253**	0.440	**3.097**	0.285	1.724
i2 7980C>T										
C	0.685	0.208	0.616	0.179	0.863	0.308	0.890	0.575	0.870	0.431
T	0.315	**4.803**	0.384	**5.601**	0.137	**3.248**	0.110	1.739	0.130	**2.319**
i5 374C>T										
C	0.554	1.431	0.599	1.951	0.582	0.878	0.625	0.926	0.740	0.549
T	0.446	0.699	0.401	0.513	0.418	1.139	0.375	1.080	0.260	1.821
e8 145A>G										
A	0.542	1.639	0.587	0.985	0.582	0.869	0.625	0.926	0.745	0.546
G	0.458	0.610	0.413	1.015	0.418	1.151	0.375	1.080	0.255	1.833
e9 1185C>T										
C	0.875	0.421	0.855	0.487	0.621	0.414	0.570	0.321	0.725	0.642
T	0.125	**2.373**	0.145	**2.054**	0.379	**2.414**	0.430	3.120	0.275	1.557
e12 1782C>T										
C	0.881	0.429	0.860	0.399	0.643	0.291	0.545	0.461	0.765	0.467
T	0.119	**2.332**	0.140	**2.505**	0.357	**3.431**	0.455	**2.168**	0.235	**2.140**
e25 3624C>T										
C	0.554	0.625	0.547	0.450	0.802	0.717	0.825	0.960	0.785	1.022
T	0.446	1.600	0.453	**2.223**	0.198	1.395	0.175	1.041	0.215	0.978
e30 4896G>A										
G	0.506	0.738	0.419	0.607	0.484	0.774	0.480	0.678	0.220	0.656
A	0.494	1.356	0.581	1.647	0.516	1.292	0.520	1.475	0.780	1.525
e30 5557A>G										
A	0.506	0.738	0.390	0.687	0.516	1.280	0.505	1.660	0.345	1.161
G	0.494	1.356	0.610	1.455	0.484	0.781	0.495	0.602	0.655	0.861
i30 6272A>G										
G	0.512	1.244	0.581	1.881	0.505	0.779	0.465	0.649	0.235	0.558
A	0.488	0.804	0.419	0.532	0.495	1.283	0.535	1.542	0.765	1.794
i30 7161G>A										
G	0.869	0.420	0.866	0.361	0.626	0.364	0.570	0.312	0.790	0.258
A	0.131	**2.379**	0.134	**2.771**	0.374	**2.749**	0.430	**3.201**	0.210	**3.874**

Table 15. *ABCC5* SNPs and their respective rEHH values. rEHH values above 2.0 were highlighted in bold and grey. (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American).

IV INFERENCES

Chapter 9 Simultaneous Genotyping of Seven Single Nucleotide Polymorphisms (SNPs) of the MDR1 gene by Single-Tube Multiplex Minisequencing

The MDR1 multidrug transporter plays an important role in protecting the body against xenobiotics. Its expression in the cells of luminal epithelium of the gut, the blood-brain barrier and blood-germ cell/fetal barrier suggest that polymorphisms that alter the function of this transporter may have an impact on the absorption, distribution and elimination of drugs, potential toxic substances and metabolites as well as contribute to susceptibility of certain diseases such as Parkinsons (Furuno et al. 2002). In fact, exon 26 3435C>T SNP in the MDR1 transporter has been associated with differences in drug side effects as well as efficacy of a variety of drugs in different individuals (Hoffmeyer et al. 2000; Hitzl et al. 2001; Kim et al. 2001; Sakaeda et al. 2001; Fellay et al. 2002; Roberts et al. 2002). However, this SNP does not result in an amino acid change, and some paradoxical associations have been observed between this SNP and MDR1 protein expression and plasma drug concentration, drug-induced side effects, and drug response, suggesting that this SNP may not be the causal SNP. Haplotype analyses of SNP 26 3435C>T and 2 other SNPs in the MDR1 gene reveal different LD block lengths in different ethnic populations raising the possibility that SNP 26 3435C>T may be linked to different unobserved causal SNPs in different study populations. It has been suggested that analyses of MDR1 SNP haplotypes rather than individual SNP genotypes would be more useful for association studies of drug response or disease susceptibility.

Thus far, the major technologies employed for genotyping SNPs in the MDR1 gene have been resequencing (Kim et al. 2001; Saito et al. 2002b), or indirect gel-based

methodologies like PCR-single strand conformation polymorphism (SSCP) (Ito et al. 2001; Kim et al. 2001; Tanabe et al. 2001) and PCR-restriction fragment length polymorphisms (RFLP) (Ameyaw et al. 2001; Cascorbi et al. 2001; Roberts et al. 2002; Tang et al. 2002), with one report of SNP 26 3435C>T genotyping using fluorogenic hybridization probes (Nauck et al. 2000). Except for the fluorogenic hybridization probe method, these methods, although effective, are relatively slow and expensive. The indirect gel-based methods of PCR-SSCP and PCR-RFLP also require subjective interpretation of gels that may contribute to an additional source of error. In addition, although the fluorogenic hybridization probe method is potentially capable of limited multiplex interrogation of 2-3 SNPs, multiplexing to interrogate more than 4 SNPs simultaneously is currently not possible with any the above methods. This report describes a method that multiplexes both the PCR as well as minisequencing steps to achieve the simultaneous genotyping of 7 SNPs across 100 kb of the MDR1 gene. This method is potentially capable of genotyping 13 or more SNPs simultaneously (Krone et al. 2002).

A major advantage of this multiplex minisequencing method of genotype determination is its cost-effectiveness. By reducing the volumes and multiplexing both the PCR and the minisequencing steps, reproducible results for the seven SNPs spanning 100 kb of the MDR1 gene could be achieved, starting with only 20 ng of genomic DNA and one-fourth of the recommended SNaPshot Mix, reducing the cost per reaction to approximately US $2 or less than 30 cents per SNP. Additionally, the multiplex PCR and minisequencing assay as described here is relatively rapid, with results for 96 DNA samples obtained within a single day.

In conclusion, the ability of this assay to rapidly and efficiently genotype 7 different SNPs across 100 kb of the MDR1 simultaneously in a single reaction should simplify

medium to high throughput multiple SNP genotyping of this gene for population-based haplotype and LD studies, as well as for studies correlating MDR1 haplotypes with drug response and/or disease. The robustness and flexibility of this genotyping assay allow it to be quickly modified for the haplotype studies of other genes as well. This thesis will demonstrate the usefulness of this genotyping assay in achieving its objective of characterizing the genetic structures of nucleotide analogue transporters *ABCC4* and *ABCC5*, two genes belonging to the same ATP-Binding Cassette superfamily as MDR1.

Chapter 10 Inference of Positive Selection on a Polymorphism in the 5' Flanking Region of the Nucleotide Analogue Transporter *ABCC4*/MRP4

Much attention has been focused on *ABCC*4, which was the first transporter shown to efflux nucleoside monophosphate analogues. Some clinical drugs are also substrates of this transporter. As a first step in establishing useful associations between variations in the *ABCC*4 gene and drug responses, the aim was to establish the haplotype and linkage disequilibrium profiles in different populations. The results from the investigation of selected polymorphisms in *ABCC*4 gene showed discrepancies of allele frequencies amongst different populations.

10.1 Variability in allele frequencies in *ABCC*4 SNPs amongst populations

Here the attention is focused on using linkage disequilibrium of haplotypes to facilitate more power in detecting association between genotype and phenotype studies of *ABCC*4. Haplotypes and LD measurements were generated from genotypes of SNP data in 5 ethnically unique populations containing 100 each of Chinese (CH), Malay (ML), Indian (IN), European American (CAU) and African American (AA). The majority of the biallelic SNPs studied were intronic. A total of 49 exonic SNPs were found in the public databases (dbSNP, TSC and JSNP as of 5th October 2002) but the reported minor allele frequencies for the non-genotyped SNPs are lower than 7.5%. The observation that allele frequencies of individual SNPs vary amongst different ethnic groups has been reported in many genes, including ABC transporters. Unlike previous findings in MDR1 and *ABCC*5 (see Chapters 8 and 11), there were no similarities between any 2 populations with regards to allele frequencies. The SNP

e11 1497C>T was only polymorphic in the African American population. The T allele was completely absent from the rest of the populations of non-African ancestry. As a potentially functionally relevant SNP, this SNP would be a good candidate for African American populations or populations with African ancestry. Similarly, allele frequencies in other SNPs including SNPs i-1 -1015G>A, i-1 -527C>G, i19 10304C>A, i20 6594G>A, and i27 7469G>A, varied between the non-African populations and the African population.

The amount of variation in allele frequencies also seemed to vary among gene loci. The variance in allele frequencies was measured using the measurement F_{st}. The range of F_{st} values for 5 populations was from 0.00 to 0.14, with a global mean of 0.052±0.041 (25 SNPs). In particular, SNPs i19 10304C>A and i20 6594G>A in *ABCC*4 had unusual high F_{st} values.

10.2 A rapidly declining LD profile of *ABCC*4

Linkage disequilibrium profile of *ABCC*4 was disperse, signified by the pale coloration in Figure 11. The suggestion that the human genome consists of 'blocks' of LD (30-100kb) interspersed with short segments of recombination hotspots might be oversimplistic. Although a segment of strong LD from SNPs i20 6594G>A and i31 1454C>T could be found, it was clearly difficult to demarcate regions of LD as 'blocks'. There were a very small number of SNP pairs in significant LD (4 out of 352 informative pairs in the African American population, using the LD statistic r^2). The pair-wise comparisons were suggestive of variable high and low levels of LD, signified by the diffuse coloration in Table. Both the half-length LD statistics (48.0-58.2 kb) and useful r^2 values (37.23-77.42 kb) in all 5 populations also confirmed this. Although *ABCC*4, *ABCC*5 and MDR1 belong to the ATP-Binding Cassette

superfamily and appear to be closely related in sharing predicted structural topology, having diverse tissue distribution and possessing the ability to transporting molecules of diverse structures out of the cell, this study clearly showed that there are genetic differences amongst these transporters. The diverse and complex genetic structure found in *ABCC*4 contrasts sharply against the *ABCC*5 and MDR1 genetic structures of strong linkage disequilibrium (Tang et al. 2002; Tang et al. 2004; Gwee et al. 2005). The half-length LD of *ABCC*4 ranges from 48.0 kb to 58.2 kb. This is comparatively shorter than both the MDR1 (Tang et al. 2002; Tang et al. 2004) and *ABCC*5 (see Chapters 8 and 11) (Gwee et al. 2005). In *ABCC*5, the entire gene studied is in a strong LD, with half-length LD of more than 144 kb while in MDR1, the half-length LD is measured to be 110 – 150 kb (Tang et al. 2002; Tang et al. 2004). It is also noted that the half-length LD of MRP1, another ABC transporter that is structurally different from *ABCC*4, *ABCC*5 or MDR1 is measured to be 23 – 48 kb (Wang et al. 2005).

10.3 Population-specific tSNPs are required

From the International Hapmap Project, validated SNP maps of well-characterized genetic markers have emerged for the majority of the human chromosomes (Ke et al. 2004) In genomic regions of high LD (e.g. haploblocks), it is possible to identify tSNPs in four global populations, in order to facilitate LD mapping of common diseases. Haplotype tagging is a means of selecting a set of non-redundant markers (tSNPs) from an initial given set of densely spaced SNPs to retain the most information about the dense map, yet reduce genotyping cost and scale. However several groups have reported that sets of tSNPs differ across different populations (Weale et al. 2003; Gwee et al. 2005). In studying specific genes that may be involved

in predisposition to complex disease or aberrant drug response, LD patterns need to be defined at a finer scale in order to obtain accuracy in candidate gene associations studies (Tishkoff and Verrelli 2003). These suggest that additional studies in specific ethnically diverse populations and individual candidate genes are still necessary. Differing sets of tSNPs were found for all 5 populations in this study of *ABCC*4 (Table 11). None of the tSNPs were shared by all the populations. Despite the general low LD exhibited, there was still a 72% reduction in number of SNPs genotyped in non-African populations and 64% reduction in African American population respectively, to achieve at least 90% coverage. Consistent with the high level of r^2 observed between the 2 SNPs i-1 -1015G>A and i-1 -527C>G as well as the stretch of genomic sequence across the 4 SNPs from i19 33474C>T to e26 3348A>G, any one of the SNPs could represent the respective regions of tight linkage and be included in disease or drug response association studies.

10.4 LRH test identifies signature of positive selection in a 5' flanking SNP of *ABCC*4

To increase power in identifying functional variants, all the polymorphisms in the *ABCC*4 gene were further screened for signatures of local positive selection. The dispersal of early humans out of the African continent infers that different selective pressures acted on human populations and positive selected genetic variants would increase the fitness (and hence survival) of specific population. Traditional tests screening for evidence of selection in candidate genes rely on degree of differentiation in polymorphic loci in different populations. Any large allele frequency differences amongst populations infer the effects of selection acting upon one or more populations but not the rest. These tests include the Wright's Fixation Index, F_{st} and

*P*excess. The Wright's Fixation Index F_{st} is essentially a measure of population subdivision and can be used to indicate the presence of selection events. The test based on *P*excess calculates the extent at which a haplotype is over-represented in a population under study as compared to an ancestral population. The average allele frequencies of the ancestral population were estimated from the African American population. Both the F_{st} and *P*excess tests are therefore dependent on the assumptions of all allele frequency differentiation measures. In addition, using allele frequency of the African American or East Asian population as the ancestral allele frequency may be erroneous, even though Bersaglieri et al., 2002 showed that their results did not defer significantly whether or not a 21% European admixture in the African American population was accounted for (Bersaglieri et al. 2004). The long range haplotype test relied on the length of linkage disequilibrium on one haplotype in relation to the frequency of that haplotype, and had been showed to be more powerful than tests dependent on measuring the degree of skewness in allele frequency distribution (Sabeti et al. 2002). Both the F_{st} and *P*excess measures on all 25 polymorphic markers were therefore utilized to obtain a smaller set of candidate polymorphisms for assessment using the more computationally intensive long range haplotype test. Despite the finding of low grade LD, the major alleles of the SNPs i-1 -1015G>A, i-1 -527C>G, and i31 1454C>T in the European American population appeared to reside in a long-ranging haplotype across the *ABCC*4 gene locus and with the exception of SNP i-1 -527C>G, showed significant departures from neutrality. The SNP i-1 -527C>G managed to show significant departures from 11 out of 12 simulated population models. The C allele of SNP i-1 -527C>G was also assessed and verified to show significant evidence of recent positive selection by using the genotype data from a panel of 30 trios from the CEPH collection (U.S. Utah residents with ancestry

from northern and western Europe) in the International Hapmap Project. This enhanced confidence that these 2 SNPs (SNPs i-1 -1015G>A and i-1 -527C>G) or other nearby SNPs within the putative promoter region and in LD with i-1 -1015G>A could be a functional variant. Using web-based transcription-binding prediction tools (listed in Materials and Methods), it was predicted that the G allele of the SNP i-1 -1015G>A resides in a binding site of the homeobox protein engrailed (*En1*), a conserved sequence abolished if an A allele is present. Localized at chromosome 2q13-q21, the human engrailed homolog 1 encodes a homeodomain-containing protein and may be important for the development of the central nervous system especially that of the mesencaphalic dopaminergic neurons (Smidt et al. 2003; Simon et al. 2004). The presence of *ABCC*4 is demonstrated in brain microvessel endothelial cells that form the blood–brain barrier (BBB) (Zhang et al. 2000) as well as the lumen of brain capillaries and in the basolateral membrane in the choroid plexus epithelium (Leggas et al. 2004). Recent studies demonstrating *ABCC*4- and *ABCC*5- mediated resistance to antiviral nucleosides suggest that one of the functions of *ABCC*4 may be involved in nucleoside transport in the blood brain barrier. Hence it would be interesting to find out if the expression of *ABCC*4 could be influenced by the SNP i-1 -1015G>A which dictates the presence of *En1* transcription factor binding site.

In summary, the strategy employed here using *ABCC*4 gene locus has identified SNP i-1 -1015G>A as displaying evidence of positive selection and uncovered a short region within the 5' flanking region as functionally important. It would be advantageous to include SNP i-1 -1015G>A in the tagging set obtained above for it could be either functional important or in linkage disequilibrium with another nearby SNP which is functional important. It is postulated that the functionally important SNP could be involved in the interindividual variability of nucleoside transport.

Chapter 11 Strong Linkage Disequilibrium at the Nucleotide Analogue Transporter *ABCC5* Gene Locus

The *ABCC5* protein transports nucleotides, cyclic-nucleotides, nucleotide analogues and certain heavy metal compounds (McAleer et al. 1999; Jedlitschky et al. 2000; Wijnholds et al. 2000; Wijnholds 2002; Wielinga et al. 2003). Akin to the function of the MDR1 transporter, *ABCC5* may play a role in regulating the bioavailability of these compounds within cells. Several clinically important antiviral and thiopurine anticancer drugs are transported by the *ABCC5* gene product (Wijnholds et al. 2000; Wijnholds 2002), and individual differences in response to these drugs may be due to cryptic differences in their gene sequences. The LD and haplotype structure of the classical ABC transporter, MDR1, was previously characterized in five different populations (Tang et al. 2002; Tang et al. 2004). A detailed haplotype and LD structure characterization of the *ABCC5* gene was now performed in these same five populations. This will enable more powerful haplotype-based association studies of functional differences compared with single SNP association analyses (Judson and Stephens 2001).

11.1 Similarities and differences in SNP, haplotype and LD profiles between populations

The *ABCC5* SNP profiles were found to be similar between the Chinese and Malays, and between the Indians and European Americans, there being no statistical difference ($P>0.05$) in observed allele frequencies at all SNP loci examined (Table 13). There was also no evidence of significant departures from evolutionary neutrality when the F_{st} statistic (Akey et al. 2002) was utilized to quantitate the variation in SNP allele

frequencies between populations as the F_{st} values for all the SNPs at the *ABCC5* gene locus were below the threshold of 0.45 (Table 1).

Similarities in haplotype profiles were also observed between the Chinese and Malays (P=0.06915); and between the European Americans and Indians (P=0.14456) (Figure 20). In contrast, the African American *ABCC5* haplotype profile differed significantly from the non-African populations (P<0.00001) (Figure 20). Consistent with the hypothesis that all modern populations are derived from a common African ancestral population, all but one of the major haplotypes (i.e. with >5% frequency) present in the non-African populations were also represented in the African American population (Figure 20). The African Americans were also observed to have the highest number of unique as well as total haplotypes, and the lowest LD of the five populations (Figures 18 and 20). Likewise, long-range LD breakdown as represented diagrammatically by the HBD (Figure 19) revealed similarities in LD decay profiles between the Chinese and Malays and between the Indians and European Americans, while the African American profile was distinct.

Thus, similar to observations made at the MDR1 gene locus (Tang et al. 2004), these current results are consistent with the "Out-of-Africa" hypothesis (Hammer 1995; Krings et al. 1997; Ingman et al. 2000) of a common non-African ancestor arising from an ancestral African population. This ancestor subsequently populated Europe / South Asia, and East / Southeast Asia.

11.2 Haplotype diversity and LD at the *ABCC5* gene locus

One notable difference between the *ABCC5* and MDR1 genes is that the former resides in a genomic region of low haplotype diversity and strong LD. Given the observed SNP allele frequencies and under an assumption of linkage equilibrium, the

expected number of haplotypes of the 16 high frequency SNPs in the non-African and African American populations should be 165 - 197 and 178, respectively. Yet only 13 - 21% and 31%, respectively, of the predicted number of haplotypes were actually observed. In contrast, the number of observed haplotypes at the MDR1 gene locus constituted 42 – 47% and 65% of the predicted number of haplotypes in the non-African and African American populations, respectively, under complete linkage equilibrium. These data suggests that haplotype diversity at the *ABCC5* gene locus is lower than that at the MDR1 gene locus. The lower number of observed haplotypes in the non-African populations supports the "Out-of-Africa" model of human evolution which suggests that the former are the descendents of small founding subgroups from Africa.

Pair-wise LD profiling using the |*D'*| statistic revealed that in all five ethnic groups, statistically significant strong LD exists between SNPs intron -1 -1205C>T and intron 30 7161G>A, which lie >100 kb apart. The half-length LD ($LD_{0.5}$) was estimated to extend beyond the size of the gene, with the African Americans displaying the shortest $LD_{0.5}$ (106 kb) followed by the Malays (144 kb), Chinese (165 kb), European Americans (266 kb) and Indians (293 kb) (Figure 18). In the MDR1 gene locus, $LD_{0.5}$ ranged from only 105 kb in the Malays to a maximum of 150 kb in European Americans (Tang et al. 2004). More striking differences were observed when the Pearson correlation (r^2) statistic was evaluated. This statistic models association power and is thus of great utility in disease association studies. When $r^2 \geqslant 0.3$ was utilized as the threshold for useful LD in association studies, useful LD extended beyond the 100 kb extent of the *ABCC5* gene (Figure 18). This observation suggests that the entire *ABCC5* gene resides within a region of strong LD and that all SNPs within the *ABCC5* gene have high associative power. This is in contrast to the 200 kb

MDR1 gene locus, where useful LD extended only from 35 kb in the Chinese and Malay populations to a maximum of 82 kb in the Indian population (Tang et al. 2004). Strong LD has also been reported in two other cancer susceptibility gene loci, ATM (140 kb) and BRCA1 (200 kb) (Bonnen et al. 2002). The observation of strong LD extending across the entire *ABCC5* gene has implications for association studies. On the one hand, it implies that fewer SNPs need to be examined in association studies. However, it also implies that once an association is made between a surrogate SNP and any functional difference, a greater region of DNA will have to be sequenced to identify the causative SNP.

11.3 Tagging SNPs at the *ABCC5* gene locus

To facilitate association studies, the "TagIT" program (Weale et al. 2003) was utilized to identify a subset of highly informative tSNPs in the *ABCC5* gene that would represent the majority of haplotypes segregating at this locus in the different populations. For each non-African population, only 4 tSNPs were sufficient to represent the majority of *ABCC5* haplotypes, with ~10% reduction in power to detect associations compared to direct assays of all 16 high minor allele frequency (>5%) SNPs (Table 14). However, the panel of 4-tSNPs that generate the maximum haplotype r^2 value in one non-African population generated lower haplotype r^2 value in other populations. Hence, common panels of 4-tSNPs for all the 4 non-African populations would result in ~10% loss in associative power (Table 14). In order to have a tSNP panel useful for all 5 populations with less than 10% loss of associative power, 6 tSNPs were necessary (Table 14). This requirement is also partly due to the fact that for the African American population, a minimum of 6 tSNPs are necessary to achieve >90% associative power. The identification of tagging SNPs is a useful tool

in association studies, potentially reducing genotyping effort by 50-70% and avoiding repeated testing of individual loci for association, which may lead to problems related to multiple testing and reduction of overall power if all ten SNPs were genotyped simultaneously. The results of 4-6 tSNPs at the *ABCC5* gene locus in the different ethnic groups is consistent with previously reported number of tSNPs required at other gene loci in other populations (Johnson et al. 2001; Weale et al. 2003).

V GENERAL DISCUSSION

Chapter 12 Concluding remarks and perspectives

12.1 *ABCC*4 and *ABCC*5: partners in crime or chalk and cheese?

There is a strong genetic component in how humans respond to drugs and xenobiotics. The interplay between drugs and the human genome (or pharmacogenomics) is therefore of immense interest in the fields of medicine and science. Single nucleotide polymorphisms, (SNPs) occurring randomly throughout the genome, may contribute to the genetic differences between individuals that influence the pharmacokinetics of drugs and hence inter-individual response to drugs. These polymorphisms affect not only metabolic enzymes such as phase I enzymes (cytochrome P450 isoenzymes) and phase II enzymes (N-acetyltransferase (NAT), glutathione S-transferase (GST) etc.) but also drug transporters as well. The MDR1 represent the classical representative of the ATP-Binding Cassette superfamily of transporters. Based on structural homology, the *ABCC*4/MRP4 and *ABCC5*/MRP5 share a P-glycoprotein-like core with MDR1 i.e. two nucleotide-binding domains (NBDs) and two membrane spanning domains (MSDs). Like MDR1, *ABCC*4 and *ABCC*5 are ubiquitously expressed in different cell types and tissues. For example, all three transporters are expressed in brain microvessel endothelial cells that form the blood–brain barrier (BBB) (Zhang et al. 2000). If polymorphisms within the MDR1 gene have been associated with variability in protein expression, plasma drug concentrations, drug-induced side effects, and drug response (Hoffmeyer et al. 2000; Hitzl et al. 2001; Kim et al. 2001; Sakaeda et al. 2001), it is reasonable to assume that polymorphisms within *ABCC*4 and *ABCC*5 may also play an equivalent role in accounting for functional differences. *ABCC*4 and *ABCC*5, the subjects of this thesis, stand out from the rest of the ABC transporters in having a common ability to confer resistance to cyclic nucleotides

adenosine 3',5'-cyclic monophosphate (cAMP) and guanosine 3',5'-cyclic monophosphate (cGMP) as well as several nucleotide analogues. These findings suggest that *ABCC*4 and *ABCC*5 as membrane transporters may be involved in the cellular homeostasis of natural nucleotides as well as having a protective function against the assault of nucleoside analogues. The initial studies demonstrating efflux of nucleoside monophosphate analogues such as 9-(2-phosphonylmethoxyethyl)adenine (PMEA), 6-mercaptopurine and thioguanine in cells overexpressing *ABCC*4 and *ABCC*5 proteins postulated that other structurally similar antiviral and antineoplastic nucleoside analogues in clinical use were likely substrates. In particular, 6-mercaptopurine and thioguanine, two thiopurine analogues of the naturally occurring bases hypoxanthine and guanine, are cytotoxic agents important in the treatment of childhood leukaemias and there is cross-resistance between them. Other therapeutic agents in clinical use have also been shown to be substrates or inhibitors of these 2 drug transporters (Table 16). These two transporters may therefore work in tandem to actively extrude these drugs out of the cell. Some publications have reported upregulation and wide interindividual variability in *ABCC*4 and *ABCC*5 mRNA and protein expression in various pathological states such as ischemic (ICM)/dilated (DCM) cardiomyopathy; small (SCLC)/non-small cell lung cancers (NSCLC); progressive familial intrahepatic cholestasis (PFIC); neuroblastoma; paediatric leukaemias and ductal pancreatic carcinoma (Sampath et al. 2002; Dazert et al. 2003; Savaraj et al. 2003; Keitel et al. 2005; Konig et al. 2005; Norris et al. 2005) with low expression in control/normal tissues. These findings have importance in the clinical treatment of leukaemias, human immunodeficiency virus 1 infection, hepatitis B viral infection, erectile dysfunction, just to name a few.

With their unique substrate profiles and ubiquitous expression, these two transporters therefore represent good candidate genes for association studies. It is not known whether the similarities between *ABCC*4 and *ABCC*5 are also extended to the genetic structures as well. This thesis therefore devised a series of studies that make use of polymorphisms to characterize the genetic profiles of these two genes for future association studies as well as compare the population genetics governing the architectures that underlie the genomic region surrounding these 2 genes.

Using a new multiplex minisequencing technique developed for the following studies, selected polymorphic markers were genotyped across the genetic loci of *ABCC*4 and *ABCC*5. Linkage disequilibrium was assessed between pairs of genetic loci using two measures of LD, *D*' and r^2. At the *ABCC*5 gene locus, the half-length LD blocks ($LD_{0.5}$) of populations of non-African ancestry were between 144 – 293 kb with that of African Americans at 106 kb (Table 17). At the *ABCC*4 gene locus, the half-length LD blocks of populations of non-African and African ancestry were half those at *ABCC*5, measuring 51 – 58 kb and 48 kb respectively. This is in spite of the fact that the genetic length of the *ABCC*4 locus (282 kb) assessed is more than twice that of *ABCC*5 locus (103 kb). Since *D*' is an indicator of recombination rates, *ABCC*5 must reside in a sequence of much lower recombination rate than *ABCC*4. When $r^2 \geqslant 0.3$ was utilized as a threshold for useful LD in association studies, useful LD was found to extend beyond the entire 100 kb of the *ABCC*5 gene locus in all five populations while it was limited to 37 - 77 kb in *ABCC*4 locus (Table 17). Associative power would therefore be greater studying SNPs in *ABCC*5 than in *ABCC*4. On the other hand, demonstration of association would refine a smaller region likely to harbor the disease or functional variant within the *ABCC*4 gene loci because linkage disequilibrium extends for shorter distances compared to the *ABCC*5 gene. Although

it is possible to use pair-wise measures such as r^2 in regions of strong LD such as *ABCC5* for association analysis, multi-locus haplotype analysis is expected to be a more powerful strategy for regions of low LD such as *ABCC4*.

From the above observations, it is apparent that *ABCC5* gene locus is in a region of strong LD while *ABCC4* gene locus has a weak LD profile. This conclusion is also extended to the haplotype profiles of the 2 transporters. In terms of number of haplotypes which were inferred from 16 SNPs in the *ABCC5* gene locus, 26 – 35 and 55 haplotypes were found in populations of non-African and African descent respectively. This indicates that a small degree of recombination has taken place. In total 130 unique *ABCC5* haplotypes were obtained. In contrast, 100 - 151 haplotypes were inferred from 25 *ABCC4* SNP markers in the non-African and African American populations respectively, with a total of 536 unique haplotypes in all populations (Table 16). While it was difficult to assess the degree of recombination between the two regions, the percentage of these observed haplotypes inferred was compared against expected haplotypes, simulated from a simulated population of the same sample size and observed allele frequency under maximum random association. Not surprisingly, the reduction in predicted haplotype number in the *ABCC4* gene locus (52.1 - 75.6%) was less aggressive than in *ABCC5* gene locus (13.2 - 30.9%). The most common haplotype in *ABCC5* accounted for 27 - 40% of total alleles in each 4 non-African population and 22.9% in the African American population while the highest haplotype frequencies of each of the 4 non-African population and African population at the *ABCC4* gene locus accounted for only 4.6 – 11.9% and 3% respectively.

Due to the greater haplotype diversity derived from the variable LD pattern in *ABCC4* gene locus, more tagging SNPs (tSNPs) were needed to account for the majority of

the diversity observed in *ABCC*4 than in *ABCC*5. Compared to 6-10 tSNPs needed for *ABCC*4, just 4-6 tSNPs were required to represent 90% of all the observed haplotypes in *ABCC*5.

These results allow one to conclude that the underlying LD and haplotype structures of *ABCC*4 and *ABCC*5 are extremely different. Indeed, current evidence indicates that the biological functions of *ABCC*4 and *ABCC*5 are probably unique to one another. *ABCC*4 gene maps to 13q32 and covers a large genomic region of 282 kb. *ABCC*5, in contrast, maps to 3q27 and its gene locus of 103 kb is less than half the size of *ABCC*4. Despite sharing a similar protein structure, the *ABCC*4 protein is only 36% identical with the *ABCC*5 protein. The relationship between *ABCC*4 and *ABCC*5 is akin to that of MDR1 and *ABCC*1/MRP1: drug transporters with considerable overlap in their substrate specificities, but retaining individual differences in substrate affinities and profiles. For example, the *ABCC*5 protein has a higher affinity for cGMP than the *ABCC*4 protein (Jedlitschky et al. 2000) while irinotecan and methotrexate are substrates of *ABCC*4 but not *ABCC*5 (Rius et al. 2003; Norris et al. 2005). The level of *ABCC*5 mRNA expression in most tissues was also shown to be much higher than *ABCC*4 (Kool et al. 1997).

Currently, there are few published studies on the evaluation of polymorphisms with regards to functional activities of *ABCC*4 and *ABCC*5 and it is likely that future studies will be directed towards using polymorphic markers to determine association to drug response, especially those of nucleotide analogues. With the tools developed through this thesis, it is hoped that more information can be gained through associating *ABCC*4/*ABCC*5 haplotypes with drug response, and thereby play a supportive role in revealing the similar-yet-different biological functions of these two transporters.

ABCC4		*ABCC5*	
Substrate	**Therapeutic Use**	**Substrate**	**Therapeutic Use**
Nonsteroidal Anti-inflammatory Drugs (NSAIDs)		*Nonsteroidal Anti-inflammatory Drugs (NSAIDs)*	
Dipyridamole	Platelet Aggregation Inhibition	Dipyridamole	Platelet Aggregation Inhibition
Indomethacin	Ankylosing Spondylitis, Bursitis, Gout, Osteoarthritis, Rheumatoid Arthritis, Shoulder Bursitis, Shoulder Tendonitis, Synovitis, Tendonitis, Tenosynovitis		
Ibuprofen	Dysmenorrhea, Fever, Juvenile Rheumatoid Arthritis, Migraine, Osteoarthritis, Pain, Rheumatoid Arthritis		
Flurbiprofen	Osteoarthritis, Rheumatoid Arthritis		
Ketoprofen	Dysmenorrhea, Osteoarthritis, Pain, Rheumatoid Arthritis		
Diclofenac	Ankylosing Spondylitis, Osteoarthritis, Rheumatoid Arthritis, Synovitis		
Celecoxib	Dysmenorrhea, Familial Adenomatous Polyposis, Osteoarthritis, Pain, Postoperative Pain, Rheumatoid Arthritis; Celebrex (Pharmacia)		
Rofecoxib	Dysmenorrhea, Familial Adenomatous Polyposis, Osteoarthritis, Pain, Postoperative Pain, Rheumatoid Arthritis; Vioxx (MSD) - Withdrawn		
Prostaglandins and Thromboxanes			
Prostaglandin PGE_1	Alprostadil, the naturally occurring prostaglandin E_1, is a vasodilating agent and a platelet-aggregation inhibitor. Erectile Dysfunction, Ductus Arteriosus-dependent Congenital Heart Disease (Neonatal)		
Estradiols, analogues and derivatives			
$E_2$17betaG	As Estradiol: Advanced Prostatic Carcinoma, Atrophic Vaginitis associated with Menopause, Female Hypogonadism, Hypoestrogenism Due To Bilateral Oophorectomy, Metastatic Breast Carcinoma, Post-Menopausal Osteoporosis Prevention, Primary Ovarian Failure, Primary Ovarian Failure, Prostatic Carcinoma, Vasomotor Symptoms associated with Menopause		
Leucovorin	Bone Marrow Suppression due to Folic Acid Antagonism, Colorectal Carcinoma, Megaloblastic Anemia, Metastatic Colorectal Cancer, Methotrexate Toxicity, Pyrimethamine Toxicity, Trimethoprim Toxicity		
Folic acid	Folate Deficiency, Folic Acid Deficient Megaloblastic Anemia		

Table 16. List of therapeutic drugs that are effluxed by *ABCC*4 or *ABCC5* and their clinical applications.

ABCC4		*ABCC5*	
Substrate	**Therapeutic Use**	**Substrate**	**Therapeutic Use**
Nucleoside analogues		*Nucleoside analogues*	
Ganciclovir	CMV Retinitis in AIDS Patients, CMV Retinitis in Immunocompromised Patients, Prevention of CMV Disease After Organ Transplant, Prevention of CMV Disease in Advanced HIV Patients		
AZT	HIV Infection, Maternal-Fetal Transmission of HIV Prevention; Retrovir (GSK)		
3TC	Chronic Hepatitis B, HIV Infection; Epivir (GSK)		
Ribavirin	Treatment of chronic hepatitis B (Schering-Plough)		
Adefovir	Chronic Hepatitis B; Hepsera (Gilead)		
		Abacavir	HIV Infection
		Cladribine	Hairy Cell Leukemia
		Gemcitabine	Non-Small Cell Lung Cancer, Pancreatic Carcinoma; Gemzar (Eli Lilly)
		Cytarabine	Acute Lymphocytic Leukemia, Acute Myeloid Leukemia, Acute Promyelocytic Leukemia, Meningeal Leukemia
		5-FU	Breast Carcinoma, Colorectal Carcinoma, Gastric Carcinoma, Metastatic Breast Carcinoma, Metastatic Colorectal Cancer, Pancreatic Carcinoma
6-Mercaptopurine (6-MP)	Acute Lymphocytic Leukemia, Acute Myeloid Leukemia, Acute Myelomonocytic Leukemia, Acute Promyelocytic Leukemia	6-Mercaptopurine (6-MP)	Acute Lymphocytic Leukemia, Acute Myeloid Leukemia, Acute Myelomonocytic Leukemia, Acute Promyelocytic Leukemia
6-Thioguanine	Acute Myeloid Leukemia (GSK)	6-Thioguanine	Acute Myeloid Leukemia (GSK)
Methotrexate (MTX)	Acute Lymphocytic Leukemia, Breast Carcinoma, Diffuse Large B-Cell Lymphoma, Head and Neck Carcinoma, Juvenile Rheumatoid Arthritis, Lung		
Uricosuric agents		*Uricosuric agents*	
Probenecid	Gouty Arthritis, Hyperuricemia, Inhibit Renal Drug Excretion	Probenecid	Gouty Arthritis, Hyperuricemia, Inhibit Renal Drug Excretion
Sulfinpyrazone	Chronic Gouty Arthritis, Hyperuricemia	Sulfinpyrazone	Chronic Gouty Arthritis, Hyperuricemia
Phosphodiesterase 5 (PDE5) Inhibitors		*Phosphodiesterase 5 (PDE5) Inhibitors*	
Sildenafil	Erectile Dysfunction; Viagra (Pfizer)	Sildenafil	Erectile Dysfunction; Viagra (Pfizer)
Topoisomerase Inhibitors			
Irinotecan	Colorectal Carcinoma, Metastatic Colorectal Cancer; Camptosar (Pfizer)		
Topotecan	Ovarian Carcinoma, Small Cell Lung		

Table 16 (continued).

Populations	*ABCC4*						*ABCC5*					
	CH	ML	IN	CAU	AA	ALL	CH	ML	IN	CAU	AA	AA
Length of gene loci examined in kb	282	282	282	282	282	282	103	103	103	103	103	103
Number of SNP loci examined for LD and haplotype analyses	25	25	25	25	25	25	16	16	16	16	16	16
Number of population samples	93	96	94	100	100	483	84	86	91	100	100	461
Linkage Disequilibrium (LD) analyses												
Half-length LD block ($LD_{0.5}$) in kb	54	51	58	58	48		144	165	266	293	160	
Useful LD ($r^2 \geq 0.3$) in kb	77.4	77.4	59.0	59.0	37.2		>103	>103	>103	>103	>103	
Haplotype analyses												
Predicted number under no recombination	26	26	26	26	26		17	17	17	17	17	
Predicted number under max recombination	185.95 ± 0.23	191.88 ± 0.34	187.61 ± 0.62	199.31 ± 0.86	199.68 ± 0.57		164.57 ± 1.63	168.93 ± 1.68	179.98 ± 1.29	197.41 ± 1.61	177.77 ± 3.53	
Observed number	104	100	116	117	151		35	28	32	26	55	
Frequency of most common haplotype in %	4.62	8.85	7.28	11.92	3.00		27.35	35.46	29.50	39.90	22.88	
tagging SNP (tSNP) sets												
Number of tSNPs with ≥90% coverage	7	7	7	7	9	10	4	4	4	4	6	6

Table 17. Summary of LD and haplotype analyses around *ABCC4* and *ABCC5* gene loci**.** (CH: Chinese; ML: Malay; IN: Indian; CAU: European American; AA: African American, ALL: all 5 populations).

12.2 So, will the real functional variant please stand up?

Many new challenges and opportunities lie in the wake of the completion of genomic sequence of the human race. With information on the position and specific sequence of every gene on any chromosome now available, it is now possible to look at a multitude of genes and investigate levels of interactions amongst sets of different genes simultaneously, instead of being limited to individual genes. The greater power of resolution has also enabled molecular biologists and geneticists to perform analysis of genome-wide scale investigations. Furthermore, as a drive to study the interindividual variation in different human populations, different polymorphisms and different transcripts of each gene can now be compared and scrutinized. As a result, the next major challenge is the analysis of these huge volumes of information in ways that allow interpretation and conclusion.

Identifying sites of functional significance

The objective of pharmacogenetics is to identify potential sites of functional significance in drug response. To help achieve this objective, this thesis outlined a series of strategies from development of a genotyping assay to linkage disequilibrium and haplotype analyses to detection of recent positive selection. The strategies were designed with a view to reduce redundancy as well as enhance power and efficiency for association studies. Association studies examine candidate marker loci or genes among affected individuals and unrelated unaffected control subjects and are therefore one of the easiest way of testing for potential sites of functional significance. The real question is how to obtain at least a suitably small subset of marker loci amongst the huge number of reported polymorphic SNPs, many of which are neutral and redundant, for association studies without losing resolution. The strategies used in this

thesis to drive the search of functional variants within two members of the human ATP-Binding Cassette (ABC) transporter superfamily of transporters, *ABCC*4 and *ABCC*5 are summarized here.

Developing a robust genotyping assay

Pharmacogenetics and genetic association studies require the analysis of multiple DNA segregating sites in large human sample pools. The work in this thesis is purely focused on profiling of single nucleotide polymorphisms (SNPs) and genetic characterization in and around candidate gene loci through the establishment of linkage disequilibrium and haplotype structures in different ethnic populations. In anticipation of such demands, the choice of having an appropriate genotyping method of sufficient throughput and affordability for a small laboratory is crucial.

The candidate gene chosen for the development of such a genotyping assay was multidrug resistance 1 (MDR1) gene, the most well-known member of the ATP-binding cassette superfamily. Encoding P-glycoprotein which has been shown to efflux a wide array of xenobiotics and drug molecules, the MDR1 gene contains several polymorphisms. Differences in SNP allele and haplotype frequencies amongst populations had been reported (Kim et al. 2001; Tang et al. 2002). Moreover, there were conflicting observations of associations linking individual SNP alleles with MDR1 expression profiles and drug concentration. A robust and rapid multiplex genotyping assay was developed to facilitate the establishment of MDR1 haplotypes (see Chapters 6 and 9). As was discussed in Chapter 1, minisequencing or SBE is a robust assay that allowed the flexibility of genotyping multiple SNP loci simultaneously. The cost effectiveness of this multiplex minisequencing technique was particularly attractive. Simultaneously amplifying multiple target sites using 5-20

ng of genomic DNA, multiplexing the minisequencing SNP detection step and reducing reaction volumes, seven MDR1 SNP loci were successfully and unambiguously genotyped (see Chapters 6 and 9). All probes used in this method were normal or HPLC purified oligos. No expensive labeled probes were needed. Allelic differentiation of minisequencing products was performed by capillary electrophoresis, on a highly-efficient and automated platform which was additionally capable of dideoxy sequencing. Further cost savings could be achieved through whole genome amplification (Hawkins et al. 2002) and DNA sample pooling. The application of this genotyping assay for haplotype analysis has been extended from MDR1 (Gwee et al. 2003; Lee et al. 2004a; Tan et al. 2004; Tang et al. 2004), to several other members of the ATP-Binding Cassette family, *ABCC*1/MRP1 (Wang et al. 2005), *ABCC*4/MRP4 (see Chapters 7 and 10), and *ABCC*5/MRP5 (see Chapters 8 and 11) (Gwee et al. 2005). This genotyping assay is uniquely positioned for future association studies of *ABCC*4 and *ABCC*5. A maximum of 10 tagging SNPs was found through LD and haplotype analyses of both genes. With the genotyping assay's ability to handle at least 10 SNPs (Tables 8 and 9), a single panel can be designed to probe these tSNPs in both case and control populations. Furthermore, many of the PCR and minisequencing primers designed for previous panels can be reused in this new panel.

Like all genotyping methods, the assay developed here is not without its limitations. Primer design is of importance. Interaction amongst multiple PCR primers should be minimized using various primer design programs available such as Vector NTI primer design program used in this thesis. Short amplicons reduce the possibility of nonspecific binding and reduce time for amplification. On the other hand, PCR of a single amplicon of relatively long size (~2 kb) is preferable to the amplification of

many short fragments. Amplicons with high GC content are best amplified separately from the amplicons with moderate or low GC content. It is, however, possible to amplify multiple GC-rich fragments in a single tube, with additives facilitating DNA denaturation. Amplification takes the most amount of time in this genotyping assay – 40 cycles of thermal cycling was required for almost all genotyping panels and random samples were monitored by gel electrophoresis. The design of the minisequencing primer is fixed by the position of the polymorphic site investigated. There are 2 directions at which the 3' end of primer probe can anneal and the one that produces less secondary structures usually gives better results. The length of the minisequencing probe can be altered by the addition of non-homologous tails which are predicted to have minimal secondary structures. The addition of poly (dGACT) tails was standardized in studies conducted in this thesis, although poly (dT), poly (dA), and poly (dC) worked just as well. The use of 5' poly (dT) tails is not recommended as they may interfere with the addition of 3' ddATP. The mobility of an oligonucleotide in capillary electrophoresis is determined by its size, nucleotide composition, and dye. While a difference of 4–6 nucleotides between minisequencing primer lengths is recommended as a guideline, it is difficult to predict the exact time at which the minisequencing products will be eluted. For an example, the genotyping results showed that the time lapse between products from short minisequencing primers of say, 20 bp and 24 bp appeared to be much quicker in the electropherogram than the time lapse between products from long minisequencing probes of 40 bp and 44 bp. As such the difference in lengths of shorter minisequencing primers were usually designed to be greater than the difference between longer minisequencing probes (Tables 8 and 9). The dye-labeled ddNTPs are of different molecular weight and therefore products coupled with different ddNTPs migrate differently. When 2

primer lengths are designed too closely and there may be a serious overlap of peaks, confusion may arise as to the identity of SNP locus especially in the case of heterozygous samples. Better results were also achieved when the melting temperatures of the longer minisequencing probes were designed to be higher than those of the shorter ones.

Capillary electrophoresis is sensitive to contaminants inside the reaction mix. When normal oligos of lengths greater than 45 bp were used, the background noise increased. Aside from cost, the preparation of minisequencing primers to be purified by HPLC took a longer time. Throughout the course of genotyping *ABCC*4 and *ABCC*5 SNPs, it was found that there was wide variability in the concentrations of minisequencing probes needed to achieve similar peak heights in the electropherogram, as the annealing efficiencies of the individual probes differed. Peak height correlated with amount of minisequencing probes used. Amounts of probes therefore had to be carefully titrated in each panel to achieve peaks of similar intensity.

While this method is potentially capable of genotyping 13 or more SNPs simultaneously (Krone et al. 2002) and this study has achieved the genotyping of 10 *ABCC*4 SNPs in a single panel (Table 9), genotyping beyond 10 SNPs in a single panel gets increasingly difficult. Optimizing multiplex PCR can be time-consuming and arduous. Moreover the addition of every SNP locus to be investigated often means re-optimizing the proportions of minisequencing probes. In this thesis, nearly 500 samples were analyzed for every genotyping panel. Any concerns of contamination from other samples, operator mistakes and false calling had to be addressed by re-sequencing random samples. Some form of automation such as the use of robotic arms could eliminate some of these concerns and increase overall efficiency but these instruments are too expensive for the average laboratory.

For a truly high throughput setting required for the simultaneous survey of even larger number of SNPs within multiple genes or across large genomic regions, there is a need to scale up from a 96-well liquid-based format to a multiplexed, microarray based platform. The use of high-density oligonucleotide arrays containing thousands of oligonucleotide tags based on the principle of minisequencing has been published in several publications (Fan et al. 2000; King et al. 2004) and the study of population-specific haplotypes using such arrays has also been reported to be successful (Raitio et al. 2001; Jain et al. 2003). While there is certainly a disparity in terms of throughput capabilities between industrial companies and small academic laboratories, the average small academic laboratory should be flexible and creative in finding innovative ways to adopt the existing large-scale genotyping methods.

Identifying a subset of SNPs useful for association studies

Association studies offer a potentially powerful approach to identify genetic variants that influence susceptibility to common disease or response to drugs. The traditional candidate gene approach has been based on attempts to detect associations between single polymorphisms within various candidate genes and phenotypes of interest. Unfortunately, many of these studies failed to find any association or are not consistently reproducible. One possible reason for the low power to detect significant associations and irreproducible results is that the approach in single polymorphism analysis adopted in such studies is oversimplistic. An alternative approach in using haplotypes or the specific combinations of alleles on an individual chromosome (Hoehe 2003) takes into account the interaction of alleles especially when the polymorphisms are in close proximity. Haplotype analysis therefore entails the simultaneous study of multiple polymorphisms within one gene, as opposed to the

study of individual SNPs. The property of alleles at two different genetic polymorphic loci having various degrees of association or linkage disequilibrium (LD) can be used to identify causative variants adjacent to a set of surrogate markers in LD if the region is associated with functional significance such as variable protein expression. Sets of SNP loci as surrogate markers in candidate gene regions are chosen such as those of *ABCC*4 and *ABCC*5 studied here can be used to narrow the location of functional variants through linkage disequilibrium analysis. In measuring the strength of linkage disequilibrium, it is important to select well-spaced SNPs with relatively high allele frequency so that these selected SNPs can be observed in the test populations used. In this thesis, the choice of surrogate SNP markers was based on the location at which they are likely to exert a functional change, reported allele frequency (if available in the public databases at point of investigation) and coverage of the whole gene region. This would ensure a greater confidence that any functional variant is likely to be within this set of selected SNPs or at least in close proximity and hence high LD with any of the markers of this set. In total, 25 and 16 polymorphic markers were chosen to cover the genetic loci of *ABCC*4 and *ABCC*5, a genomic distance of 282 kb and 103 kb respectively. From the genotype allele frequencies, haplotypes were estimated using EM algorithm and pair-wise LD measures amongst SNPs were calculated. When haplotypes within a region of strong linkage disequilibrium are inferred, a large proportion of the diversity of haplotypes can be explained by a few common haplotypes and many others with rare haplotype frequencies. As such the majority of the diversity within a region can be captured by only genotyping SNPs shared amongst the common haplotypes. In contrast, a region of low linkage disequilibrium implies that there are numerous haplotypes of low frequencies and will therefore require more tagging SNPs (tSNPs).

The juxtaposition of contrasting genetic profiles is clearly demonstrated here in this thesis. *ABCC*4 gene locus lies in a genomic region marked by rapidly decaying LD amongst SNPs and large numbers of haplotypes while *ABCC*5 gene locus is in a region of strong linkage disequilibrium characterized with few haplotypes. These findings have strong implications for haplotype association studies involving these 2 drug transporter genes. A genomic sequence or gene locus in investigation such as *ABCC*5 that has a strong LD pattern would require very few surrogate marker loci to represent the region, but it requires additional finer mapping to look for the functional variant. This proved to be true in the LD and haplotype analyses of *ABCC*5 as intensive pair-wise LD between *ABCC*5 SNPs meant that only 4-6 tagging SNPs were required to adequately describe the majority of the haplotype diversity with the gene locus. Genotyping fewer tagging SNPs for *ABCC*5 compared to *ABCC*4 can reduce overall time and costs of association studies. In the event that association between any of these tagging SNPs and complex trait is to be found, then the genomic region surrounding this surrogate marker that has to be re-sequenced for the presence of the real functional variant in LD with the marker locus is likely to be large. These additional as-yet-unknown sequence variants will require further investigation. On the other hand, *ABCC*4 with a weak LD pattern requires more surrogate marker loci to describe adequately the genomic region, but the functional variant should be in close proximity to any marker loci that may show association with a phenotypic change. The rapidly declining linkage disequilibrium profile of *ABCC*4 required more tagging SNPs than *ABCC*5 to explain 90% of the complexity within the gene locus. However, this implied that one only needs to search within a shorter genomic region for a functional variant if any association with disease / drug phenotype was to be observed as compared to the strong LD profile of *ABCC*5 gene locus.

Identifying functional variants though signatures of positive selection

A lot of effort has been focused at detecting the presence of positive selection during the evolution of a gene (Fay et al. 2001; Akey et al. 2004; Altshuler and Clark 2005). With the increasing amount of DNA polymorphism data in human and primate populations, there is a renewed interest in assessing the neutral theory of evolution. Genes such as those of drug enzymes and transporters may confer an evolutionary advantage for the organism or certain human populations to adapt to a different environment or physiological requirement. Therefore it can be reasonably assumed that the drive to attain evolutionary 'fitness' in these genes may be positive selection. As past selective events may affect the flanking haplotypes of the functional or disease variant, the Long-Range Haplotype test (Sabeti et al. 2002) was modified to analyze SNPs within the candidate genes. Five human populations were subjected to assessment of positive selection to identify population-specific SNPs which may be crucial in dictating evolutionary events at the *ABCC*4 and *ABCC*5 transporter loci. Using the inherent LD underlying a long-ranging haplotype, a polymorphic marker SNP i-1 -1015G>A within the 5' flanking region of *ABCC*4 gene locus was successfully identified, showing significant departures from neutrality. This 5' flanking region could potentially be within the promoter of *ABCC*4 and of functional significance. Importantly the observation of high haplotype homozygosity (2.31) in the A allele of SNP -1 -1015G>A was limited to the European American population but not four other populations studied (Chinese, Malay, Indian, and African American). When subjected to the LRH test, this allele in the European American population demonstrated significant departure from neutrality under all four different population models and three different recombination rate assumptions ($P<0.05$). Just

as importantly, there was also supportive evidence of recent positive selection using an independent population of similar ethnic background. Genotype data were obtained from 30 sets of CEPH trios (Utah residents with ancestry from northern and western Europe) and it was hypothesized that some evidence of a long-ranging haplotype surrounding the A allele of SNP -1 -1015G>A could be detected in nearby SNPs. This effort uncovered a SNP that was located 488 bp away and had high associative power with SNP -1 -1015G>A (r^2 = 0.92 for European American population). The G allele of the SNP rs869951 (SNP -1 -527C>G) possessed a high rEHH value and also showed significant departures from neutrality. It can be proposed that this SNP is under the influence of genetic hitchhiking from the SNP -1 -1015G>A. When selective forces acted on the A allele of SNP -1 -1015G>A, the adjacent G allele of SNP rs869951 (SNP -1 -527C>G) was in tight linkage disequilibrium with this allele and there was insufficient time for recombination. Hence evidence of recent positive selection could be demonstrated for this short stretch of genomic region found within the putative promoter of *ABCC*4.

The web-based transcription binding tool MatInspector [http://www.genomatix.de/products/MatInspector/index.html] predicted that the G allele of the SNP i-1 -1015G>A resided in a putative binding site of homeobox protein engrailed (*En1*), a conserved sequence that would be abolished if the alternative A allele was present. Regarded as developmental genes, the engrailed-1 (*En1*) and its other paralog *En2* are expressed in the central nervous system from early in development to adulthood (Smidt et al. 2003; Simon et al. 2004). They are involved in the pattern formation of the mid/hindbrain as well as neurogenesis, neuronal differentiation, regulation of apoptosis of mesencephalic dopaminergic (mesDA)

neurons (Alberi et al. 2004; Simon et al. 2004). They are therefore implicated in neurological disorders such as Parkinson's disease.

Given this, it is not inconceivable that the transcription factor *En1* may directly influence the expression of the *ABCC*4 in the brain and brain capillaries, thereby regulating the uptake of neurotoxic xenobiotics and maintaining the integrity of the central nervous system. This may also suggest a mechanism for the etiology of neurological disorders, interindividual variability in drug responses and adverse events of clinically relevant drugs that are *ABCC*4 substrates.

This polymorphism SNP -1 -1015G>A within the putative promoter region could therefore be of functional importance and warrants further investigation. Inclusion of this SNP within the tSNP set of European American population established in these studies might also gain strength in looking for the true functional variant.

12.3 Sources of complexity to LD and haplotype studies

There are several major contributors that may add to the complexity in studying linkage equilibrium and haplotype patterns.

Population diversity

While most populations share common SNP variants and haplotype patterns inherited from the common ancestor population, frequencies of these SNP variants and haplotype patterns may be similar or different amongst populations. The greater extensive LD and lower number of observed haplotypes in the non-African populations at both *ABCC*4 and *ABCC*5 supports the "Out-of-Africa" model of human evolution which suggests that the former are the descendents of small founding subgroups from Africa. With greater haplotype diversity, more tSNPs were needed to provide coverage in the African American populations at both loci. Comparing amongst the populations of non-African descent was not as simple. As was discussed in *ABC*C5, using Fisher's exact test and pair-wise *F*st tests, the Chinese and Malay populations as well as the Indian and European American populations typically showed more similarities in allele frequencies between them. This similarity was also extended to haplotypes inferred and their haplotype frequencies. This trend was not replicated at the *ABCC*4 gene locus where it seemed that each of the non-African population possessed its specific set of haplotypes with varying frequencies. Moreover, while similar sets of tSNPs could be found for all non-African populations, each population had its own best performing set of tSNPs i.e. with the highest haplotype r^2.

The results in this thesis underscore to need to study haplotypes in specific populations. In regions of low haplotype diversity and high LD such as that surrounding the *ABCC*5 gene locus where common haplotypes representing diverse

human populations are expressed, one could extrapolate findings to another population. For example, the data in Malay population closely parallels that in the Chinese population so it might seem a good idea to base any results in association studies with *ABCC5* in one population to the other. All *ABCC5* haplotypes in populations of non-African descent could be found in the African American population, alluding to the possibility of using African American haplotypes to predict genotypes of other ethnic background (Evans et al. 2004). On the other hand, very few similar haplotypes were shared across populations at the *ABCC4* gene locus. Predictably, using the haplotype information from the African American population to map functional variants for another population would be inaccurate. Therefore at a region of variable LD and high haplotype diversity such as that of *ABCC4*, more efforts to include a high-definition LD mapping are needed for each studied population in order to obtain any power in association studies.

Genotyping errors and deviation of Hardy-Weinberg equilibrium

Another issue concerning the application of linkage disequilibrium (LD) and haplotype analysis is genotyping errors. Genotyping errors may produce Type I results especially if there is a systematic error in calling one genotype over another. In familial studies, genotyping errors can appear as Mendelian inconsistency and these cases of Mendelian inconsistency can be resolved by either retyping or setting as missing data. However such error checks cannot be performed in unrelated individuals. To minimize this error, the genotypes should be ensured to be in Hardy-Weinberg proportions and sample sizes should not be too small. On the other hand, SNPs showing deviation from Hardy-Weinberg should not be discarded immediately. In cases the violation from HWE is expected for variants close to a susceptibility

locus. Nevertheless, the genotyping protocol has to be reviewed and SNPs deviating from HWE should preferably be genotyped using a different method to ensure correct calling. In this thesis the allele and genotype frequency distributions of several polymorphic SNPs deviated from HWE prior to Bonferonni correction for multiple-locus testing. Population substructure was ruled out as there was no consistent deviation from HWE of genotype data in a single population. Nevertheless all minisequencing electropherograms were rechecked and recalled independently. Sequencing random samples as well as samples in which genotypes were too difficult to call did not show that there were gross inaccuracies in the genotype data initially obtained.

Additionally this thesis also followed the recommendation in implementing haplotype inference programs such as those based on Expectation-Maximization algorithm (Arlequin, SNPHAP and TAGIT) that were not as sensitive to deviations of HWE for the genetic characterization of *ABCC*4 and *ABCC*5 (Niu et al. 2002; Niu 2004) even though it must be stated that HWE is one of the assumptions of EM.

12.4 Future perspectives

There are many polymorphisms besides nonsynonymous SNPs which are potentially functionally important. Polymorphisms may introduce stop codons and cause a premature truncation of protein or be located in splice sites may lead to aberrant splicing or cause a complete frameshift (Crawford et al. 2005). Alternatively spliced variants have been reported for both *ABCC*4 and *ABCC*5. Lamba et al detected an *ABCC*4 cDNA that contained two additional exons within intron 1. The inclusion of either one or both of these 2 exons produced a frame shift and a premature termination codon (PTC). Of the three major variants, one of them with an additional single exon

1b was a predominant transcript. Lamba proposed that these conserved PTC exons may facilitate translational re-initiation and lead to protein diversity (Lamba et al. 2003).

A short transcript, SMRP had been reported by Suzuki et al (Suzuki et al. 1997; Suzuki et al. 2000). First cloned from a cisplatin-resistant human lung adenocarcinoma cell line, this transcript corresponds to the 3' end of the full-length transcript of *ABCC5* and translates to a protein of only 946 amino acids (instead of 1437 amino acids from the full transcript of *ABCC5*/MRP5) and Northern blot analysis showed its expression in various tissues (Suzuki et al. 1997). Low levels of SMRP mRNA had been successfully detected in human auricular and ventricular tissue samples by real-time PCR (Dazert et al. 2003). Significant increase of this transcript was also observed in samples of patients suffering from ischemic cardiomyopathy (ICM) over levels in samples from patients with dilated cardiomyopathy (DCM) or nonfailing heart tissues. Increased levels of expression of both *ABCC5* and SMRP were detected by real-time PCR in a human nonsmall cell lung cancer line after induction of adriamycin and increased mRNA expression was also observed in adriamycin-resistant cell lines using RNase protection assay and Northern blotting analysis (Yoshida et al. 2001). In another report using Taqman real-time PCR, no significant correlation could be found between expression of both MRP and SMRP in 53 children with de novo acute myeloid leukemia (AML) with overall survival or remission rate (Steinbach et al. 2003a). Likewise, this same technique also showed no significant association of both *ABCC5* and SMRP in 103 children diagnosed with previously untreated acute lymphoblastic leukemia (ALL) with poor response to chemotherapy and poor prognosis (Steinbach et al. 2003b). It is unknown whether other splice variant forms of *ABCC5* exist.

Some of these spliced forms of *ABCC*4 and *ABCC*5 may give rise to truncated proteins that retain the reading frame but have lost parts of functional domains. There are two challenges to this area of study. One is to detect the presence of these alternate splice products and investigate whether these proteins are functional or affect normal function of the two nucleotide analogue transporters. The second challenge is to elucidate if the polymorphisms can affect the splicing mechanism of *ABCC*4 and *ABCC*5. Intronic and synonymous SNPs can be potentially functionally important and influence either splicing accuracy or efficiency (Cartegni et al. 2002). For example, it was found that a translationally silent synonymous SNP can result in inefficient inclusion of an exon leading to an unstable, inactive protein (Cartegni and Krainer 2002). It would be of interest to find out if a polymorphism within *ABCC*4 or *ABCC*5 causes the alteration of correct splicing patterns and thereby introduces phenotypic variability.

Polymorphisms that reside within the regulatory regions of the genes are also important. SNPs in the promoter can alter the expression of the gene by affecting transcription binding while SNPs in the untranslated regions (5'UTR and 3'UTR) can affect mRNA stability, translation or transport. In the putative promoter region of *ABCC*4, this thesis identified one SNP -1 -1015G>A within a long range haplotype that is under the influence of recent positive selection and alters a putative binding site of a transcription factor. In the putative promoter region of *ABCC*5, three novel low-frequency SNPs were discovered. The promoters of both the *ABCC*4 and *ABCC*5 genes are not yet characterized and this presents a unique opportunity for research.

12.5 Conclusions

In the context of candidate gene analysis, this thesis describes a model in which a series of strategies designed to facilitate future genetic association studies can be adopted. Instead of conventional analysis using single polymorphic variants, the objective was to harness information from the dependencies amongst SNPs using haplotype association analysis. The establishment of a multiplex minisequencing genotyping technique allowed the rapid and robust investigation of multiple polymorphic markers around the loci of two nucleotide analogue transporters *ABCC*4 and *ABCC*5. These two genes encode for transporters demonstrated to efflux nucleotide analogues useful for antiretroviral therapy and chemotherapy as well as other clinically relevant drug molecules. By carefully selecting polymorphic markers based on distance and functionality, the LD and haplotype architectures around *ABCC*4 and *ABCC5* gene loci were further characterized in five different populations, including Chinese, Malay, Indian, European American and African American. The two loci exhibited contrasting genetic structures; whereas LD decayed rapidly across 282 kb of *ABCC*4, it remained constantly strong across 103 kb of *ABCC5* gene. A modest number of population-specific tagging SNP sets that best described the underlying haplotype structure useful for these genes were also defined. These results also demonstrate the capabilities of the algorithms chosen to analyze large number of genotype data and across large genomic distances.

Transporter genes such as *ABCC*4 and *ABCC5* that are involved in xenobiotic efflux may be under selective pressures as the migration of early modern humans out of Africa to other continents exposed them to new environments, new pathogens and new diets. One of the signatures of recent positive selection is the occurrence of a haplotype with long range LD given its frequency. The final strategy to detect

evidence of recent positive selection singularized the allele A of SNP i-1 -1015G>A within the putative promoter region of *ABCC*4 in an European American population, but not other populations. The discovery that selective forces occur in transporter genes that may affect drug response *in vivo* such as MDR1 (Tang et al. 2004), *ABCC*1/*MRP* (Wang et al. 2005) and *ABCC*4/*MRP*4 (Chapters 7 and 10) is also important for the study of human evolution and adaptation.

The flexible model described in this thesis can easily be adopted for haplotype association studies of other candidate genes in addition to those involved in drug response or human diseases and would therefore augment efforts in disclosing the functional variants within these genes.

VI REFERENCES

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191-197

Adachi M, Reid G, Schuetz JD (2002) Therapeutic and biological importance of getting nucleotides out of cells: a case for the ABC transporters, MRP4 and 5. Adv Drug Deliv Rev 54:1333-1342

Adkins RM (2004) Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. BMC Genet 5:22

Akashi H (1999) Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221-238

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805-1814

Alberi L, Sgado P, Simon HH (2004) Engrailed genes are cell-autonomously required to prevent apoptosis in mesencephalic dopaminergic neurons. Development 131:3229-3236

Altshuler D, Clark AG (2005) Genetics. Harvesting medical information from the human family tree. Science 307:1052-1053

Ameyaw MM, Regateiro F, Li T, Liu X, Tariq M, Mobarek A, Thornton N, Folayan GO, Githang'a J, Indalo A, Ofori-Adjei D, Price-Evans DA, McLeod HL (2001) MDR1 pharmacogenetics: frequency of the C3435T mutation in exon 26 is significantly influenced by ethnicity. Pharmacogenetics 11:217-221

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309.

Bader JS (2001) The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics 2:11-24

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99-111

Barnes MR (2002) SNP and mutation data on the Web - hidden treasures for uncovering. Comparative and Functional Genomics 3:67-74

Belinsky MG, Bain LJ, Balsara BB, Testa JR, Kruh GD (1998) Characterization of MOAT-C and MOAT-D, new members of the MRP/cMOAT subfamily of transporter proteins. J Natl Cancer Inst 90:1735-1741.

Berezowski V, Landry C, Dehouck MP, Cecchelli R, Fenart L (2004) Contribution of glial cells and pericytes to the mRNA profiles of P-glycoprotein and multidrug resistance-associated proteins in an in vitro model of the blood-brain barrier. Brain Res 1018:1-9

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111-1120

Bodo A, Bakos E, Szeri F, Varadi A, Sarkadi B (2003) The role of multidrug transporters in drug availability, metabolism and toxicity. Toxicol Lett 140-141:133-143.

Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL (2002) Haplotype and linkage disequilibrium architecture for human cancer- associated genes. Genome Res 12:1846-1853.

Borst P, Balzarini J, Ono N, Reid G, de Vries H, Wielinga P, Wijnholds J, Zelcer N (2004) The potential impact of drug transporters on nucleoside-analog-based antiviral chemotherapy. Antiviral Res 62:1-7

Borst P, Elferink RO (2002) Mammalian ABC transporters in health and disease. Annu Rev Biochem 71:537-592

Brookfield JF (2003) Human prehistory: the message from linkage disequilibrium. Curr Biol 13:R86-87

Bush JA, Li G (2002) Cancer chemoresistance: the relationship between p53 and multidrug transporters. Int J Cancer 98:323-330

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231-238.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106-120

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285-298

Cartegni L, Krainer AR (2002) Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nat Genet 30:377-384

Cascorbi I, Gerloff T, Johne A, Meisel C, Hoffmeyer S, Schwab M, Schaeffeler E, Eichelbaum M, Brinkmann U, Roots I (2001) Frequency of single nucleotide polymorphisms in the P-glycoprotein drug transporter MDR1 gene in white subjects. Clin Pharmacol Ther 69:169-174. t&artType=abs&id=a114164&target=

Chen X, Sullivan PF (2003) Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. Pharmacogenomics J 3:77-96

Chen ZS, Lee K, Kruh GD (2001) Transport of cyclic nucleotides and estradiol 17-beta-D-glucuronide by multidrug resistance protein 4. Resistance to 6-mercaptopurine and 6-thioguanine. J Biol Chem 276:33747-33754

Chen ZS, Lee K, Walther S, Raftogianis RB, Kuwano M, Zeng H, Kruh GD (2002) Analysis of methotrexate and folate transport by multidrug resistance protein 4 (ABCC4): MRP4 is a component of the methotrexate efflux system. Cancer Res 62:3144-3150

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-122

Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev 13:296-302

Crawford DC, Akey DT, Nickerson DA (2005) The Patterns of Natural Variation in Human Genes. Annu Rev Genomics Hum Genet

Crawford DC, Nickerson DA (2005) Definition and clinical importance of haplotypes. Annu Rev Med 56:303-320

Dallas S, Schlichter L, Bendayan R (2004) Multidrug resistance protein (MRP) 4- and MRP 5-mediated efflux of 9-(2-phosphonylmethoxyethyl)adenine by microglia. J Pharmacol Exp Ther 309:1221-1229

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232.

Dantzig AH, Pratt SE, Shepard RL (2003) MRP5 confers resistance to 5-FU and transports a phosphorylated metabolite of 5-FU. Proc Am Assoc Cancer Res 44:735

Davidson JD, Ma L, Iverson PW, Lesoon A, Jin S, Horwitz L, Gallery M, Slapak CA (2002a) Human multidrug resistance protein 5 (MRP5) confers resistance to gemcitabine. Proc Am Assoc Cancer Res 43:780-781

Davidson JD, Ma L, Iverson PW, Lesoon A, Jin S, Horwitz L, Gallery M, Slapak CA (2002b) Human multi-drug resistance protein 5 (MRP5) confers resistance to gemcitabine. Proc Am Assoc Cancer Res 43:3868

Davidson S (2000) Research suggests importance of haplotypes over SNPs. Nat Biotechnol 18:1134-1135

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544-548

Dazert P, Meissner K, Vogelgesang S, Heydrich B, Eckel L, Bohm M, Warzok R, Kerb R, Brinkmann U, Schaeffeler E, Schwab M, Cascorbi I, Jedlitschky G, Kroemer HK (2003) Expression and localization of the multidrug resistance protein 5 (MRP5/ABCC5), a cellular export pump for cyclic nucleotides, in human heart. Am J Pathol 163:1567-1577

Dean M, Rzhetsky A, Allikmets R (2001) The human ATP-binding cassette (ABC) transporter superfamily. Genome Res 11:1156-1166.

Deloukas P, Bentley D (2004) The HapMap project and its application to genetic studies of drug response. Pharmacogenomics J 4:88-90

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci U S A 97:10483-10488.

Efferth T (2001) The human ATP-binding cassette transporter genes: from the bench to the bedside. Curr Mol Med 1:45-65

Evans DM, Cardon LR, Morris AP (2004) Genotype prediction using a dense map of SNPs. Genet Epidemiol 27:375-384

Evans WE, McLeod HL (2003) Pharmacogenomics--drug disposition, drug targets, and side effects. N Engl J Med 348:538-549

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947-959

Fan JB, Chen X, Halushka MK, Berno A, Huang X, Ryder T, Lipshutz RJ, Lockhart DJ, Chakravarti A (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. Genome Res 10:853-860

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227-1234

Fellay J, Marzolini C, Meaden ER, Back DJ, Buclin T, Chave JP, Decosterd LA, Furrer H, Opravil M, Pantaleo G, Retelska D, Ruiz L, Schinkel AH, Vernazza

P, Eap CB, Telenti A (2002) Response to antiretroviral treatment in HIV-1-infected individuals with allelic variants of the multidrug resistance transporter 1: a pharmacogenetics study. Lancet 359:30-36

Furuno T, Landi MT, Ceroni M, Caporaso N, Bernucci I, Nappi G, Martignoni E, Schaeffeler E, Eichelbaum M, Schwab M, Zanger UM (2002) Expression polymorphism of the blood-brain barrier component P-glycoprotein (MDR1) in relation to Parkinson's disease. Pharmacogenetics 12:529-534

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109-111.

Goldstein DB, Ahmadi KR, Weale ME, Wood NW (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. Trends Genet 19:615-622

Goldstein DB, Weale ME (2001) Population genomics: linkage disequilibrium holds the key. Curr Biol 11:R576-579.

Gottesman MM, Ambudkar SV (2001) Overview: ABC transporters and human disease. J Bioenerg Biomembr 33:453-458

Gottesman MM, Fojo T, Bates SE (2002) Multidrug resistance in cancer: role of ATP-dependent transporters. Nature Rev Cancer 2:48-58.

Grabe N (2002) AliBaba2: context specific identification of transcription factor binding sites. In Silico Biol 2:S1-15

Graff CL, Pollack GM (2004) Drug transport at the blood-brain barrier and the choroid plexus. Curr Drug Metab 5:95-108

Gray IC, Campbell DA, Spurr NK (2000) Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 9:2403-2408

Guo Y, Kotova E, Chen ZS, Lee K, Hopper-Borge E, Belinsky MG, Kruh GD (2003) MRP8, ATP-binding cassette C11 (ABCC11), is a cyclic nucleotide efflux pump and a resistance factor for fluoropyrimidines 2',3'-dideoxycytidine and 9'-(2'-phosphonylmethoxyethyl)adenine. J Biol Chem 278:29509-29514

Gwee PC, Tang K, Chua JM, Lee EJ, Chong SS, Lee CG (2003) Simultaneous genotyping of seven single-nucleotide polymorphisms in the MDR1 gene by single-tube multiplex minisequencing. Clin Chem 49:672-676.

Gwee PC, Tang K, Sew PH, Lee EJ, Chong SS, Lee CG (2005) Strong linkage disequilibrium at the nucleotide analogue transporter ABCC5 gene locus. Pharmacogenet Genomics 15:91-104

Halldorsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S (2004a) Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Res 14:1633-1640

Halldorsson BV, Istrail S, De La Vega FM (2004b) Optimal selection of SNP markers for disease association studies. Hum Hered 58:190-202

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376-378

Hawkins TL, Detter JC, Richardson PM (2002) Whole genome amplification--applications and advances. Curr Opin Biotechnol 13:65-67

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409-411

Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH (1997) Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques 23:504-511

Hirrlinger J, Konig J, Dringen R (2002) Expression of mRNAs of multidrug resistance proteins (Mrps) in cultured rat astrocytes, oligodendrocytes, microglial cells and neurones. J Neurochem 82:716-719.

Hitzl M, Drescher S, van der Kuip H, Schaffeler E, Fischer J, Schwab M, Eichelbaum M, Fromm MF (2001) The C3435T mutation in the human MDR1 gene is associated with altered efflux of the P-glycoprotein substrate rhodamine 123 from CD56+ natural killer cells. Pharmacogenetics 11:293-298

Hoehe MR (2003) Haplotypes and the systematic analysis of genetic variation in genes and genomes. Pharmacogenomics 4:547-570

Hoffmeyer S, Burk O, von Richter O, Arnold HP, Brockmoller J, Johne A, Cascorbi I, Gerloff T, Roots I, Eichelbaum M, Brinkmann U (2000) Functional polymorphisms of the human multidrug-resistance gene: multiple sequence variations and correlation of one allele with P- glycoprotein expression and activity in vivo. Proc Natl Acad Sci U S A 97:3473-3478.

Hsu TM, Law SM, Duan S, Neri BP, Kwok PY (2001) Genotyping single-nucleotide polymorphisms by the invader assay with dual-color fluorescence polarization detection. Clin Chem 47:1373-1377

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713

Issa JP, Kantarjian HM, Kirkpatrick P (2005) Azacitidine. Nat Rev Drug Discov 4:275-276

Ito S, Ieiri I, Tanabe M, Suzuki A, Higuchi S, Otsubo K (2001) Polymorphism of the ABC transporter genes, MDR1, MRP1 and MRP2/cMOAT, in healthy Japanese subjects. Pharmacogenetics 11:175-184.

Jain M, Thorstenson YR, Faulkner DM, Pourmand N, Jones T, Au M, Oefner PJ, White KP, Davis RW (2003) Genotyping African haplotypes in ATM using a co-spotted single-base extension assay. Hum Mutat 22:214-221

Jedlitschky G, Burchell B, Keppler D (2000) The multidrug resistance protein 5 functions as an ATP-dependent export pump for cyclic nucleotides. J Biol Chem 275:30069-30074.

Jedlitschky G, Tirschmann K, Lubenow LE, Nieuwenhuis HK, Akkerman JW, Greinacher A, Kroemer HK (2004) The nucleotide transporter MRP4 (ABCC4) is highly expressed in human platelets and present in dense granules, indicating a role in mediator storage. Blood 104:3603-3610

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233-237

Jorajuria S, Dereuddre-Bosquet N, Naissant-Storck K, Dormont D, Clayette P (2004) Differential expression levels of MRP1, MRP4, and MRP5 in response to human immunodeficiency virus infection in human macrophages. Antimicrob Agents Chemother 48:1889-1891

Judson R, Stephens JC (2001) Notes from the SNP vs. haplotype front. Pharmacogenomics 2:7-10.

Juliano RL, Ling V (1976) A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants. Biochim Biophys Acta 455:152-162

Kaminskas E, Farrell A, Abraham S, Baird A, Hsieh LS, Lee SL, Leighton JK, Patel H, Rahman A, Sridhara R, Wang YC, Pazdur R (2005a) Approval summary: azacitidine for treatment of myelodysplastic syndrome subtypes. Clin Cancer Res 11:3604-3608

Kaminskas E, Farrell AT, Wang YC, Sridhara R, Pazdur R (2005b) FDA drug approval summary: azacitidine (5-azacytidine, Vidaza) for injectable suspension. Oncologist 10:176-182

Ke X, Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287-288

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577-588

Keitel V, Burdelski M, Warskulat U, Kuhlkamp T, Keppler D, Haussinger D, Kubitz R (2005) Expression and localization of hepatobiliary transport proteins in progressive familial intrahepatic cholestasis. Hepatology 41:1160-1172

Kim RB, Leake BF, Choo EF, Dresser GK, Kubba SV, Schwarz UI, Taylor A, Xie HG, McKinsey J, Zhou S, Lan LB, Schuetz JD, Schuetz EG, Wilkinson GR (2001) Identification of functionally variant MDR1 alleles among European Americans and African Americans. Clin Pharmacol Ther 70:189-199. t&artType=abs&id=a117412&target=

King GC, Di Giusto DA, Wlassoff WA, Giesebrecht S, Flening E, Tyrelle GD (2004) Proofreading genotyping assays and electrochemical detection of SNPs. Hum Mutat 23:420-425

Klein I, Sarkadi B, Varadi A (1999) An inventory of the human ABC proteins. Biochim Biophys Acta 1461:237-262.

Klokouzas A, Wu CP, van Veen HW, Barrand MA, Hladky SB (2003) cGMP and glutathione-conjugate transport in human erythrocytes. Eur J Biochem 270:3696-3708

Konig J, Hartel M, Nies AT, Martignoni ME, Guo J, Buchler MW, Friess H, Keppler D (2005) Expression and localization of human multidrug resistance protein (ABCC) family members in pancreatic carcinoma. Int J Cancer 115:359-367

Kool M, de Haas M, Scheffer GL, Scheper RJ, van Eijk MJ, Juijn JA, Baas F, Borst P (1997) Analysis of expression of cMOAT (MRP2), MRP3, MRP4, and MRP5, homologues of the multidrug resistance-associated protein gene (MRP1), in human cancer cell lines. Cancer Res 57:3537-3547

Kool M, van der Linden M, de Haas M, Baas F, Borst P (1999) Expression of human MRP6, a homologue of the multidrug resistance protein gene MRP1, in tissues and cancer cells. Cancer Res 59:175-182

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet 1:539-559

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S (1997) Neandertal DNA sequences and the origin of modern humans. Cell 90:19-30

Krone N, Braun A, Weinert S, Peter M, Roscher AA, Partsch CJ, Sippell WG (2002) Multiplex minisequencing of the 21-hydroxylase gene as a rapid strategy to confirm congenital adrenal hyperplasia. Clin Chem 48:818-825.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139-144

Kruh GD, Belinsky MG (2003) The MRP family of drug efflux pumps. Oncogene 22:7537-7552

Kwok PY (2001) Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet 2:235-258

Lai L, Tan TM (2002) Role of glutathione in the multidrug resistance protein 4 (MRP4/ABCC4)-mediated efflux of cAMP and resistance to purine analogues. Biochem J 361:497-503

Lamba JK, Adachi M, Sun D, Tammur J, Schuetz EG, Allikmets R, Schuetz JD (2003) Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. Hum Mol Genet 12:99-109

Lee CG, Tang K, Cheung YB, Wong LP, Tan C, Shen H, Zhao Y, Pavanni R, Lee EJ, Wong MC, Chong SS, Tan EK (2004a) MDR1, the blood-brain barrier transporter, is associated with Parkinson's disease in ethnic Chinese. J Med Genet 41:e60

Lee K, Belinsky MG, Bell DW, Testa JR, Kruh GD (1998) Isolation of MOAT-B, a widely expressed multidrug resistance-associated protein/canalicular multispecific organic anion transporter-related transporter. Cancer Res 58:2741-2747

Lee K, Klein-Szanto AJ, Kruh GD (2000) Analysis of the MRP4 drug resistance profile in transfected NIH3T3 cells. J Natl Cancer Inst 92:1934-1940

Lee SH, Walker DR, Cregan PB, Boerma HR (2004b) Comparison of four flow cytometric SNP detection assays and their use in plant improvement. Theor Appl Genet 110:167-174

Lee VH (2000) Membrane transporters. Eur J Pharm Sci 11 Suppl 2:S41-50

Leggas M, Adachi M, Scheffer GL, Sun D, Wielinga P, Du G, Mercer KE, Zhuang Y, Panetta JC, Johnston B, Scheper RJ, Stewart CF, Schuetz JD (2004) Mrp4 confers resistance to topotecan and protects the brain from chemotherapy. Mol Cell Biol 24:7612-7621

Lewontin RC (1974) Annotation: the analysis of variance and the analysis of causes. Am J Hum Genet 26:400-411.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195

Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. Am J Hum Genet 71:1129-1137

Lockhart AC, Tirona RG, Kim RB (2003) Pharmacogenetics of ATP-binding cassette transporters in cancer and chemotherapy. Mol Cancer Ther 2:685-698

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799-810

Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, Davenport R, Miller RD, Kwok PY (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? Nat Genet 27:371-372

Marzolini C, Paus E, Buclin T, Kim RB (2004) Polymorphisms in human MDR1 (P-glycoprotein): recent advances and clinical relevance. Clin Pharmacol Ther 75:13-33

McAleer MA, Breen MA, White NL, Matthews N (1999) pABC11 (also known as MOAT-C and MRP5), a member of the ABC family of proteins, has anion transporter activity but does not confer multidrug resistance when overexpressed in human embryonic kidney 293 cells. J Biol Chem 274:23541-23548.

Meyer UA (2004) Pharmacogenetics - five decades of therapeutic lessons from genetic diversity. Nat Rev Genet 5:669-676

Meyer Zu Schwabedissen HE, Grube M, Heydrich B, Linnemann K, Fusch C, Kroemer HK, Jedlitschky G (2005) Expression, localization, and function of MRP5 (ABCC5), a transporter for cyclic nucleotides, in human placenta and cultured human trophoblasts: effects of gestational age and cellular differentiation. Am J Pathol 166:39-48

Min DI, Ellingrod VL (2002) C3435T mutation in exon 26 of the human MDR1 gene and cyclosporine pharmacokinetics in healthy subjects. Ther Drug Monit 24:400-404

Mueller JC, Andreoli C (2004) Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity. Bioinformatics 20:786-787

Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev 12:657-663

Nauck M, Stein U, von Karger S, Marz W, Wieland H (2000) Rapid detection of the C3435T polymorphism of multidrug resistance gene 1 using fluorogenic hybridization probes. Clin Chem 46:1995-1997

Nies AT, Spring H, Thon WF, Keppler D, Jedlitschky G (2002) Immunolocalization of multidrug resistance protein 5 in the human genitourinary system. J Urol 167:2271-2275.

Nikaido H (2002) How are the ABC transporters energized? Proc Natl Acad Sci U S A 99:9609-9610

Niu T (2004) Algorithms for inferring haplotypes. Genet Epidemiol 27:334-347

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157-169

Norris MD, Smith J, Tanabe K, Tobin P, Flemming C, Scheffer GL, Wielinga P, Cohn SL, London WB, Marshall GM, Allen JD, Haber M (2005) Expression of multidrug transporter MRP4/ABCC4 is a marker of poor prognosis in neuroblastoma and confers resistance to irinotecan in vitro. Mol Cancer Ther 4:547-553

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839-1851

Pascolo L, Fernetti C, Pirulli D, Crovella S, Amoroso A, Tiribelli C (2003) Effects of maturation on RNA transcription and protein expression of four MRP genes in human placenta and in BeWo cells. Biochem Biophys Res Commun 303:259-265

Pati N, Schowinsky V, Kokanovic O, Magnuson V, Ghosh S (2004) A comparison between SNaPshot, pyrosequencing, and biplex invader SNP genotyping methods: accuracy, cost, and throughput. J Biochem Biophys Methods 60:1-12

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.[see comment]. Science 294:1719-1723

Pratt S, Shepard RL, Kandasamy RA, Johnston PA, Perry W, 3rd, Dantzig AH (2005) The multidrug resistance protein 5 (ABCC5) confers resistance to 5-fluorouracil and transports its monophosphorylated metabolites. Mol Cancer Ther 4:855-863

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1-14

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179-1189

Qin ZS, Niu T, Liu JS (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 71:1242-1247

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23:4878-4884

Raitio M, Lindroos K, Laukkanen M, Pastinen T, Sistonen P, Sajantila A, Syvanen AC (2001) Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. Genome Res 11:471-482

Rebbeck TR, Spitz M, Wu X (2004) Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet 5:589-597

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204.

Reid G, Wielinga P, Zelcer N, De Haas M, Van Deemter L, Wijnholds J, Balzarini J, Borst P (2003a) Characterization of the transport of nucleoside analog drugs by the human multidrug resistance proteins MRP4 and MRP5. Mol Pharmacol 63:1094-1103

Reid G, Wielinga P, Zelcer N, van der Heijden I, Kuil A, de Haas M, Wijnholds J, Borst P (2003b) The human multidrug resistance protein MRP4 functions as a prostaglandin efflux transporter and is inhibited by nonsteroidal antiinflammatory drugs. Proc Natl Acad Sci U S A 100:9244-9249

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, et al. (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223-228

Ritter CA, Jedlitschky G, Meyer zu Schwabedissen H, Grube M, Kock K, Kroemer HK (2005) Cellular export of drugs and signaling molecules by the ATP-binding cassette transporters MRP4 (ABCC4) and MRP5 (ABCC5). Drug Metab Rev 37:253-278

Rius M, Nies AT, Hummel-Eisenbeiss J, Jedlitschky G, Keppler D (2003) Cotransport of reduced glutathione with bile salts by MRP4 (ABCC4) localized to the basolateral hepatocyte membrane. Hepatology 38:374-384

Roberts RL, Joyce PR, Mulder RT, Begg EJ, Kennedy MA (2002) A common P-glycoprotein polymorphism is associated with nortriptyline-induced postural hypotension in patients treated for major depression. Pharmacogenomics J 2:191-196

Robertson A (1975a) Gene frequency distributions as a test of selective neutrality. Genetics 81:775-785

Robertson A (1975b) Letters to the editors: Remarks on the Lewontin-Krakauer test. Genetics 80:396

Sabatti C, Risch N (2002) Homozygosity and linkage disequilibrium. Genetics 160:1707-1719

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting

recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Saito S, Iida A, Sekine A, Miura Y, Ogawa C, Kawauchi S, Higuchi S, Nakamura Y (2002a) Identification of 779 genetic variations in eight genes encoding members of the ATP-binding cassette, subfamily C (ABCC/MRP/CFTR. J Hum Genet 47:147-171

Saito S, Iida A, Sekine A, Miura Y, Ogawa C, Kawauchi S, Higuchi S, Nakamura Y (2002b) Three hundred twenty-six genetic variations in genes encoding nine members of ATP-binding cassette, subfamily B (ABCB/MDR/TAP), in the Japanese population. J Hum Genet 47:38-50

Sakaeda T, Nakamura T, Horinouchi M, Kakumoto M, Ohmoto N, Sakai T, Morita Y, Tamura T, Aoyama N, Hirai M, Kasuga M, Okumura K (2001) MDR1 genotype-related pharmacokinetics of digoxin after single oral administration in healthy Japanese subjects. Pharm Res 18:1400-1404

Sampath J, Adachi M, Hatse S, Naesens L, Balzarini J, Flatley RM, Matherly LH, Schuetz JD (2002) Role of MRP4 and MRP5 in biology and chemotherapy. AAPS PharmSci 4:E14

Savaraj N, Wu C, Wangpaichitr M, Kuo MT, Lampidis T, Robles C, Furst AJ, Feun L (2003) Overexpression of mutated MRP4 in cisplatin resistant small cell lung cancer cell line: collateral sensitivity to azidothymidine. Int J Oncol 23:173-179

Schinkel AH, Jonker JW (2003) Mammalian drug efflux transporters of the ATP binding cassette (ABC) family: an overview. Adv Drug Deliv Rev 55:3-29

Schuetz JD, Connelly MC, Sun D, Paibir SG, Flynn PM, Srinivas RV, Kumar A, Fridland A (1999) MRP4: A previously unidentified factor in resistance to nucleoside-based antiviral drugs. Nat Med 5:1048-1051

Schug J, Overton GC (1997) Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. Proc Int Conf Intell Syst Mol Biol 5:268-271

Schulze TG, Chen YS, Akula N, Hennessy K, Badner JA, McInnis MG, DePaulo JR, Schumacher J, Cichon S, Propping P, Maier W, Rietschel M, Nothen MM, McMahon FJ (2002) Can long-range microsatellite data be used to predict short-range linkage disequilibrium? Hum Mol Genet 11:1363-1372

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF (2003) Minimal haplotype tagging. Proc Natl Acad Sci U S A 100:9900-9905

Sham P, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA Pooling: a tool for large-scale association studies. Nat Rev Genet 3:862-871

Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet 12:771-776.

Simon HH, Thuret S, Alberi L (2004) Midbrain dopaminergic neurons: control of their cell fate by the engrailed transcription factors. Cell Tissue Res 318:53-61

Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 76:377-383.

Smeets PH, van Aubel RA, Wouterse AC, van den Heuvel JJ, Russel FG (2004) Contribution of multidrug resistance protein 2 (MRP2/ABCC2) to the renal excretion of p-aminohippurate (PAH) and identification of MRP4 (ABCC4) as a novel PAH transporter. J Am Soc Nephrol 15:2828-2835

Smidt MP, Smits SM, Burbach JP (2003) Molecular mechanisms underlying midbrain dopamine neuron development and function. Eur J Pharmacol 480:75-88

Steinbach D, Lengemann J, Voigt A, Hermann J, Zintl F, Sauerbrey A (2003a) Response to chemotherapy and expression of the genes encoding the multidrug resistance-associated proteins MRP2, MRP3, MRP4, MRP5, and SMRP in childhood acute myeloid leukemia. Clin Cancer Res 9:1083-1086

Steinbach D, Wittig S, Cario G, Viehmann S, Mueller A, Gruhn B, Haefer R, Zintl F, Sauerbrey A (2003b) The multidrug resistance-associated protein 3 (MRP3) is associated with a poor outcome in childhood ALL and may account for the worse prognosis in male patients and T-cell immunophenotype. Blood 102:4493-4498

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC (2003) Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 55:27-36

Suzuki T, Nishio K, Sasaki H, Kurokawa H, Saito-Ohara F, Ikeuchi T, Tanabe S, Terada M, Saijo N (1997) cDNA cloning of a short type of multidrug resistance protein homologue, SMRP, from a human lung cancer cell line. Biochem Biophys Res Commun 238:790-794

Suzuki T, Sasaki H, Kuh HJ, Agui M, Tatsumi Y, Tanabe S, Terada M, Saijo N, Nishio K (2000) Detailed structural analysis on both human MRP5 and mouse mrp5 transcripts. Gene 242:167-173

Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930-942

Takasawa K, Terasaki T, Suzuki H, Sugiyama Y (1997) In vivo evidence for carrier-mediated efflux transport of 3'-azido-3'-deoxythymidine and 2',3'-dideoxyinosine across the blood-brain barrier via a probenecid-sensitive transport system. J Pharmacol Exp Ther 281:369-375

Tan EK, Drozdzik M, Bialecka M, Honczarenko K, Klodowska-Duda G, Teo YY, Tang K, Wong LP, Chong SS, Tan C, Yew K, Zhao Y, Lee CG (2004) Analysis of MDR1 haplotypes in Parkinson's disease in a white population. Neurosci Lett 372:240-244

Tanabe M, Ieiri I, Nagata N, Inoue K, Ito S, Kanamori Y, Takahashi M, Kurata Y, Kigawa J, Higuchi S, Terakawa N, Otsubo K (2001) Expression of P-glycoprotein in human placenta: relation to genetic polymorphism of the multidrug resistance (MDR)-1 gene. J Pharmacol Exp Ther 297:1137-1143.

Tang K, Ngoi SM, Gwee PC, Chua JM, Lee EJ, Chong SS, Lee CG (2002) Distinct haplotype profiles and strong linkage disequilibrium at the MDR1 multidrug transporter gene locus in three ethnic Asian populations. Pharmacogenetics 12:437-450.

Tang K, Wong LP, Lee EJ, Chong SS, Lee CG (2004) Genomic evidence for recent positive selection at the human MDR1 gene locus. Hum Mol Genet

The International Hapmap Consortium (2003) The International HapMap Project. Nature 426:789-796

Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293-340

van Aubel RA, Smeets PH, Peters JG, Bindels RJ, Russel FG (2002) The MRP4/ABCC4 gene encodes a novel apical organic anion transporter in

human kidney proximal tubules: putative efflux pump for urinary cAMP and cGMP. J Am Soc Nephrol 13:595-603

van Aubel RA, Smeets PH, van den Heuvel JJ, Russel FG (2005) Human organic anion transporter MRP4 (ABCC4) is an efflux pump for the purine end metabolite urate with multiple allosteric substrate binding sites. Am J Physiol Renal Physiol 288:F327-333

von Ahsen N, Richter M, Grupp C, Ringe B, Oellerich M, Armstrong VW (2001) No influence of the MDR-1 C3435T polymorphism or a CYP3A4 promoter polymorphism (CYP3A4-V allele) on dose-adjusted cyclosporin A trough concentrations or rejection incidence in stable renal transplant recipients. Clin Chem 47:1048-1052

Wang L, Xu Y (2003) Haplotype inference by maximum parsimony. Bioinformatics 19:1773-1780

Wang Z, Wang B, Tang K, Lee EJ, Chong SS, Lee CG (2005) A functional polymorphism within the MRP1 gene locus identified through its genomic signature of positive selection. Hum Mol Genet 14:2075-2087

Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73:551-565

Wielinga PR, Reid G, Challa EE, van der Heijden I, van Deemter L, de Haas M, Mol C, Kuil AJ, Groeneveld E, Schuetz JD, Brouwer C, De Abreu RA, Wijnholds J, Beijnen JH, Borst P (2002) Thiopurine metabolism and identification of the thiopurine metabolites transported by MRP4 and MRP5 overexpressed in human embryonic kidney cells. Mol Pharmacol 62:1321-1331

Wielinga PR, van der Heijden I, Reid G, Beijnen JH, Wijnholds J, Borst P (2003) Characterization of the MRP4- and MRP5-mediated transport of cyclic nucleotides from intact cells. J Biol Chem 278:17664-17671

Wijnholds J (2002) Drug resistance caused by multidrug resistance-associated proteins. Novartis Found Symp 243:69-79

Wijnholds J, Mol CA, van Deemter L, de Haas M, Scheffer GL, Baas F, Beijnen JH, Scheper RJ, Hatse S, De Clercq E, Balzarini J, Borst P (2000) Multidrug-resistance protein 5 is a multispecific organic anion transporter able to transport nucleotide analogs. Proc Natl Acad Sci U S A 97:7476-7481.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A 102:7882-7887

Wu CP, Woodcock H, Hladky SB, Barrand MA (2005) cGMP (guanosine 3',5'-cyclic monophosphate) transport across human erythrocyte membranes. Biochem Pharmacol 69:1257-1262

Xu H, Wu X, Spitz MR, Shete S (2004) Comparison of haplotype inference methods using genotypic data from unrelated individuals. Hum Hered 58:63-68

Yoshida M, Suzuki T, Komiya T, Hatashita E, Nishio K, Kazuhiko N, Fukuoka M (2001) Induction of MRP5 and SMRP mRNA by adriamycin exposure and its overexpression in human lung cancer cells resistant to adriamycin. Int J Cancer 94:432-437

Young LC, Campling BG, Voskoglou-Nomikos T, Cole SP, Deeley RG, Gerlach JH (1999) Expression of multidrug resistance protein-related genes in lung cancer: correlation with drug response. Clin Cancer Res 5:673-680

Zelcer N, Reid G, Wielinga P, Kuil A, van der Heijden I, Schuetz JD, Borst P (2003) Steroid and bile acid conjugates are substrates of human multidrug-resistance protein (MRP) 4 (ATP-binding cassette C4). Biochem J 371:361-367

Zhang J, Vingron M, Hoehe MR (2005) Haplotype reconstruction for diploid populations. Hum Hered 59:144-156

Zhang K, Calabrese P, Nordborg M, Sun F (2002) Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71:1386-1394

Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19:1300-1301

Zhang Y, Han H, Elmquist WF, Miller DW (2000) Expression of various multidrug resistance-associated protein (MRP) homologues in brain microvessel endothelial cells. Brain Res 876:148-153

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89-100