Am. J. Hum. Genet. 75:190–203, 2004

Large-Scale Single-Nucleotide Polymorphism (SNP) and Haplotype Analyses, Using Dense SNP Maps, of 199 Drug-Related Genes in 752 Subjects: the Analysis of the Association between Uncommon SNPs within Haplotype Blocks and the Haplotypes Constructed with Haplotype-Tagging SNPs

Naoyuki Kamatani,[1] Akihiro Sekine,[2] Takuya Kitamoto,[2] Aritoshi Iida,[2] Susumu Saito,[2] Akifumi Kogame,[1] Eisuke Inoue,[1] Manabu Kawamoto,[1] Masayoshi Harigai,[1] and Yusuke Nakamura[2]

[1]Division of Genomic Medicine, Department of Advanced Biomedical Engineering and Science, and Institute of Rheumatology, Tokyo Women's Medical University, Tokyo; and [2]Laboratory of Genotyping, RIKEN SNP Research Center, Riken Yokohama Institute, Yokohama

To optimize the strategies for population-based pharmacogenetic studies, we extensively analyzed single-nucleotide polymorphisms (SNPs) and haplotypes in 199 drug-related genes, through use of 4,190 SNPs in 752 control subjects. Drug-related genes, like other genes, have a haplotype-block structure, and a few haplotype-tagging SNPs (htSNPs) could represent most of the major haplotypes constructed with common SNPs in a block. Because our data included 860 uncommon (frequency <0.1) SNPs with frequencies that were accurately estimated, we analyzed the relationship between haplotypes and uncommon SNPs within the blocks (549 SNPs). We inferred haplotype frequencies through use of the data from all htSNPs and one of the uncommon SNPs within a block and calculated four joint probabilities for the haplotypes. We show that, irrespective of the minor-allele frequency of an uncommon SNP, the majority (mean ± SD frequency 0.943 ± 0.117) of the minor alleles were assigned to a single haplotype tagged by htSNPs if the uncommon SNP was within the block. These results support the hypothesis that recombinations occur only infrequently within blocks. The proportion of a single haplotype tagged by htSNPs to which the minor alleles of an uncommon SNP were assigned was positively correlated with the minor-allele frequency when the frequency was <0.03 ($P < .000001$; $n = 233$ [Spearman's rank correlation coefficient]). The results of simulation studies suggested that haplotype analysis using htSNPs may be useful in the detection of uncommon SNPs associated with phenotypes if the frequencies of the SNPs are higher in affected than in control populations, the SNPs are within the blocks, and the frequencies of the SNPs are >0.03.

Introduction

Responses to drugs vary from subject to subject. The symptoms in some patients may improve dramatically during treatment with a given drug, whereas other patients develop severe adverse reactions to the same compound. The differences in the response to a compound between individuals is of huge importance (Lazarou et al. 1998). The individual differences in drug response are, at least in part, attributable to genetic information, although the effects of other factors, such as sex, age, and disease, are not negligible (Evans and Relling 1999; Meyer 2000; Evans and McLeod 2003).

Studies of the association between genetic variation and drug response have entered a new era. Recent rapid advances in human genome research have enabled researchers to genotype at numerous polymorphic loci, such as SNP loci, within a short period of time (Ohnishi et al. 2001; Jurinke et al. 2002). Moreover, ~4,000,000 of the estimated 10,000,000 common SNPs are already known (Smigielski et al. 2000; Kruglyak and Nickerson 2001; Hirakawa et al. 2002).

Which polymorphic loci in which genes should be examined? Once a drug is administered, it is absorbed and distributed to its site of action, where it interacts with targets, undergoes metabolism, and is then excreted. Each of these processes could potentially involve an individual difference in drug response. The genes of various enzymes responsible for the metabolism of the exogenous compounds, transporters, and drug targets should be examined. It is now becoming clear that vir-

Received December 19, 2003; accepted for publication May 10, 2004; electronically published June 16, 2004.

Address for correspondence and reprints: Dr. Naoyuki Kamatani, Division of Statistical Genetics, Institute of Rheumatology, Tokyo Women's Medical University, 10-22 Kawada-cho, Shinjuku-ku, Tokyo 162-0054, Japan. E-mail: kamatani@ior.twmu.ac.jp

0002-9297/2004/7502-0005$15.00

tually every pathway of drug metabolism will eventually be found to have genetic variations (Meyer 2000).

When the data from many loci on the human genome are analyzed to associate genetic variations with phenotypes, several aspects of population genetics are required in addition to those of molecular genetics. Recent studies of data from many individuals at many SNP loci have clarified that the human genome has a haplotype (or linkage disequilibrium [LD]) block structure (Clark et al. 1998; Collins et al. 1999; Kruglyak 1999). Within a block, LD is strong, and the number of major haplotypes is limited. Genes related to specific drugs are likely to have a similar haplotype block structure. Knowledge about this structure is available; recent studies have elucidated that haplotype or diplotype configuration (haplotype combination), rather than SNP genotype, is often the principal determinant of phenotypic consequences (Horikawa et al. 2000; Hugot et al. 2001; Judson and Stephens 2001; Ogura et al. 2001; Rioux et al. 2001). Indeed, haplotype analysis has been useful in predicting the outcome of drug therapy for a β_2-adrenergic stimulator (Drysdale et al. 2000), methotrexate (Urano et al. 2002), and sulfasalazine (Tanaka et al. 2002).

In the present investigation, we genotyped DNA from 752 individuals, for a total of 4,190 SNPs in 199 genes that code for enzymes of drug metabolism and transport. On the basis of the knowledge obtained by this analysis, we aimed to establish the optimal strategy for such population-based pharmacogenetic studies. Our questions included which SNPs should be genotyped, whether haplotype-based methods are useful, how to detect phenotype-associated uncommon SNPs, and what sample size is necessary for the detection of significance. During the analysis, we found that the haplotype block structure is useful not only for common SNPs but also for uncommon SNPs within the blocks. This is because the majority of the minor alleles of each uncommon SNP are assigned to a single major haplotype. We extensively analyzed these assignments, because this knowledge and the technology to analyze it are likely to be useful in various population-based genetic studies.

Material and Methods

Subjects

The present study was approved by the genome ethics committee of Tokyo Women's Medical University and by that of the Pharma SNP Consortium. The subjects from whom DNA was obtained were recruited from volunteers. Informed consent was obtained from each of the subjects. A total of 1,032 volunteers were recruited, and DNA samples from 752 subjects randomly selected from among these volunteers were used for the present study. Among the 752 subjects, 449 were male and 303 were female. The mean ± SD ages of the subjects were 36.1 ± 11.5 years for the men and 40.6 ± 11.3 years for the women.

Genotyping

The Invader assay combines structure-specific cleavage enzymes and a universal fluorescent resonance energy transfer system (Ohnishi et al. 2001). Allele-specific oligonucleotide pairs and invasive probes were designed and supplied by Third Wave Technologies. The procedures for identifying the SNP genotypes have been published elsewhere (Kwiatkowski et al. 1999).

Construction of Haplotype Blocks

We developed our own computer program to construct haplotype blocks; however, the algorithm we implemented was essentially based on methods described elsewhere (Zhu et al. 2003). A block was constructed using two steps. In the first step, an initial interval within which all pairwise D' values were ≥ 0.9 was constructed. When haplotypes were inferred within this interval, there were a few major haplotypes whose frequencies were $\geq 5\%$. The combined frequency of the major haplotypes was $\geq 90\%$. To this initial interval, an adjacent SNP was added, and the inference of the haplotypes was performed using the adjacent SNP in addition to the SNPs within the interval. If any additional major haplotype did not appear by this inference, then the new SNP was added to the other SNPs to make a new interval. This procedure was repeated in each of the 5′ and 3′ directions until the inclusion of an adjacent SNP generated an additional major haplotype. The resulting interval spanned all the SNPs examined except for two, the inclusion of which increased the number of major haplotypes until they were defined as a haplotype block. In addition to the above method for haplotype block construction, we also used the method of Gabriel et al. (2002) for comparison.

The inference of the population frequencies of the haplotypes was performed by use of the expectation-maximization (EM) algorithm (Excoffier and Slatkin 1995; Long et al. 1995; Kitamura et al. 2002). However, this method is limited, in that it can handle only ≤ 20 loci. Thus, haplotype-tagging SNPs (htSNPs) were used when >20 loci were involved. In this case, the inference of the haplotypes was performed by partition-ligation–EM (Qin et al. 2002), and the number of the SNPs was reduced by selecting htSNPs that the EM algorithm could handle.

Inference of Haplotype Frequencies and Selection of htSNPs

The inference of haplotypes within a haplotype block was performed using only the SNPs with minor-allele frequencies ≥ 0.1. For the inference, the EM algorithm (Excoffier and Slatkin 1995; Long et al. 1995) was used with LDSUPPORT software (Kitamura et al. 2002).

The htSNPs were selected using major haplotypes that explained either $\geq 95\%$ or $\geq 90\%$ of all the haplotypes within a block. The methods used to select htSNPs were essentially the same as those described elsewhere (Daly et al. 2001; Johnson et al. 2001; Patil et al. 2001). Specifically, we used the phase 2 method of Avi-Itzhak et al. (2003). The phase 1 method of Avi-Itzhak et al. (2003) was also used to compare results.

Analysis of the Relationship between Uncommon SNPs and Major Haplotypes

To describe the relationship between uncommon SNPs and haplotypes, we had to define a probability space. Imagine that there is a haplotype block in which a total of N linked polymorphic loci are present. Let H_i denote the ith complete haplotype within the block. Here, a complete haplotype is defined as a list of N alleles at all N loci (one allele per locus). The sample space Ω is defined as a set of all complete haplotypes in all subjects in the population. Here, H_i is redefined as a subset of Ω whose elements are H_i. Let X denote the minor allele at one of N loci within the block. X can be redefined as a set of all complete haplotypes with X in the lists. Then, the redefined X becomes a subset of Ω. Haplotypes defined only by htSNPs can be interpreted as incomplete haplotypes. Thus, an incomplete haplotype A_i is defined as the ith list of alleles at only htSNP loci within the block. A_i is redefined as a set of complete haplotypes consistent with A_i at htSNP loci. When the above definitions are used, a complete haplotype H_i, an incomplete haplotype A_i, and X, the minor alleles of an SNP are all defined as subsets of Ω; they are interpreted as events on the same probability space. The complements of those events, $\bar{X}$ and $\bar{A}_i$, can also be defined. Note that those concepts are not easily defined or stated otherwise.

The association between an uncommon SNP and the haplotypes tagged by htSNPs within a block was analyzed as follows. The inference of the haplotypes was performed by using the genotypes at all htSNPs and one of the uncommon SNPs (frequency <0.1) within a block. The haplotype inference was performed using either LDSUPPORT (Kitamura et al. 2002) or PHASE (Stephens et al. 2001) software. LDSUPPORT infers, on the basis of the EM algorithm, both the population haplotype frequencies and the diplotype distribution (the posterior distribution of diplotype configuration) of each subject, whereas PHASE uses the Markov chain–Monte Carlo method and the coalescence model for the inference of the haplotypes. After the haplotype inference, the following joint probabilities were obtained for all A_i: $P(A_i,X)$, $P(\bar{A}_i,X)$, $P(A_i,\bar{X})$, and $P(\bar{A}_i,\bar{X})$. Using such estimated probabilities, we could calculate $P(A_i \mid X) = P(A_i,X)/[P(A_i,X) + P(\bar{A}_i,X)]$. A_j, the incomplete haplotype that maximizes $P(A_i \mid X)$, was then selected as the incomplete haplotype to which X was assigned. $P(X \mid A_j) = P(A_j,X)/[P(A_j,X) + P(A_j,\bar{X})]$ and $P(A_j \mid X) = P(A_j,X)/[P(A_j,X) + P(\bar{A}_j,X)]$ were calculated as the measures for the assignment of an uncommon SNP to the incomplete haplotypes. The detailed procedure to analyze the assignment of an uncommon SNP to the incomplete haplotypes will be described in the "Results" section, using a block as an example.

Simulation to Detect Phenotype-Associated Uncommon SNPs with Haplotypes Tagged by htSNPs

The probability of detecting significance through use of A_j instead of X in association studies depends on factors such as $P(X \mid \psi)$, $P(X \mid \bar{\psi})$, $P(A_j \mid \bar{X})/P(A_j \mid X)$, and M_1,M_2, the numbers of affected and control subjects, where ψ and $\bar{\psi}$ denote the set of complete haplotypes in affected subjects and its complement, respectively. The algorithm for the simulation is described in appendix A.

Results

Genes and SNPs Included in the Study

We genotyped DNA from 752 Japanese individuals at 4,190 SNPs in 199 genes that are either drug-related genes or are candidates for being drug-related genes. The list of the genes, their chromosomal locations, and the number of SNPs in each gene that we studied is presented as a Web supplement (Genstat Web site). Among them, 4,153 SNP loci in 193 genes were on autosomes, and 37 SNPs in 6 genes were X linked. Some of the genes are known to be associated clinically with drug reactions, whereas others code for transporters, oxidoreductases, various transferases, and miscellaneous proteins.

Accuracy of the Data

The SNPs genotyped in the present study were derived from the results of an SNP discovery in which DNA from 48 subjects was used (Iida et al. 2001*a*, 2001*b*, 2002*a*, 2002*b*, 2003; Saito et al. 2002). Among all the SNPs found by this discovery, 4,190 SNPs that were successfully subjected to the genotyping procedure, as described in the "Material and Methods" section, were used to genotype DNA from 752 subjects. Hardy-Weinberg equilibrium was always checked, and data that deviated strongly from the equilibrium were either submitted to retyping or discarded.

For the present study, accuracy of the results was ab-

solutely necessary, because we constructed haplotypes on the basis of the accurate genotypes at many loci. In addition, the accuracy of the genotype data was vital to assigning the minor alleles of each uncommon SNP to major haplotypes. Therefore, we first evaluated the accuracy of the genotype data. The frequency of typing errors was empirically 0.0001045, according to data obtained by retyping the same materials. However, we reevaluated the rate of the mistyping as follows, because retyping can lead to the same errors, given the nature of the materials. The 752 samples included those from 449 men. These samples should not have had any heterozygous loci at X-linked SNPs. The researchers who performed the genotyping were kept totally blinded to the sex of each subject from whom the DNA was obtained. There were 37 X-linked SNP loci. When the missing data were excluded, there were 16,479 X-linked genotypes obtained from men, and, among them, 12 were heterozygous. Therefore, the frequency of the mistyping from the homozygous to the heterozygous state was estimated to be 0.00073.

The mistyping rate from the heterozygous to the homozygous state was evaluated as follows. The minor-allele frequency was estimated for each X-linked locus, by use of only the data from the apparently homozygous genotypes from the men. Note that the heterozygous genotypes at such loci are false and should be excluded. From those data, the expected proportion of the homozygous genotypes at X-linked loci for women was estimated to be 0.9139. On the other hand, the observed proportion of homozygous genotypes at X-linked loci for women was 0.9169. The difference between the observed and expected proportions was only 0.0033. We have already shown that there were few mistypings from the homozygous to the heterozygous state (~0.00073). Therefore, if the mistyping rate from the heterozygous to the homozygous state was high, the observed proportion of the homozygous genotypes at the female X-linked loci would be significantly higher than the expected proportion. Thus, although the genotyping error rates may be still underestimated, the data from both the retyping experiments and the X-linked loci indicate that the mistyping rates were very small (<0.001).

Distribution of the Frequencies of SNPs in Autosomes

Although the SNPs included in the present study were discovered in 96 chromosomes, a considerable number of SNPs were present at proportions <.01 among the 1,504 autosomal chromosomes. Thus, 898 (21.6%) of the autosomal SNPs that we examined were present in percentages <10% among the 1,504 chromosomes. Figure 1 shows the histogram of the minor-allele frequency for all 4,153 autosomal SNPs in 1,504 chromosomes. This is as expected, because even the SNPs with very low frequencies have a chance to be included in a small number of samples, given that the sampling is a stochastic process.

Construction of Haplotype Blocks

A haplotype, or LD block, is defined as a sizable region over which there is little evidence for historical recombination and within which only a few major haplotypes are observed. Because the definition of a haplotype block is somewhat ambiguous, there are some discrepancies between the intervals of the blocks constructed by different researchers, even when they should span the same region. In fact, we have constructed haplotype blocks through use of methods described by Zhu et al. (2003) and Gabriel et al. (2002). The haplotype blocks constructed using these two different methods were a little different, as will be discussed below. Still, the concept of the block has many benefits and is expected to be useful for extracting important information about the region, using data from association studies. For the construction of haplotype blocks, SNPs with minor-allele frequencies >0.1 or >0.2 are usually used. However, the frequencies of SNPs estimated using a small number of subjects are quite unreliable. In the present study, however, the frequencies were rather reliable, because they were estimated using a large number of subjects (752) from the same homogeneous population.

Among 4,153 autosomal SNPs, only 4,104 SNPs unambiguously mapped to the chromosomal regions were used for the construction of haplotype blocks. Among them, the minor-allele frequencies were ⩾0.1 in 3,244 SNPs (common SNPs), whereas they were <0.1 in the remaining 860 SNPs (uncommon SNPs) (fig. 2). There are no unequivocal definitions for "common" and "uncommon" SNPs; however, common and uncommon (or rare) SNPs denote, throughout the present article, SNPs with minor-allele frequencies of ⩾0.1 and <0.1, respectively.

Using the 3,244 common SNPs, we identified a total of 356 blocks in 179 autosomal genes. Among the 3,244 common SNPs, 3,132 (96.6%) SNPs were within the blocks, whereas the remaining 112 (3.5%) SNPs were out of the blocks (fig. 2). Within the blocks, independent measures of pairwise LD did not decrease substantially with distance (data not shown). The mean ± SD number of blocks in an autosomal gene was 1.93 ± 8.75 (median 1). Some blocks were much larger than others.

Lengths of Blocks and the Regions between Blocks

Among the 356 blocks in autosomes, either the 5′ or 3′ ends of the blocks were the same as the edges of the regions spanning the sets of examined SNPs in 182 blocks. The apparent ends of such blocks may or may not be the real ends, because blocks may extend farther.

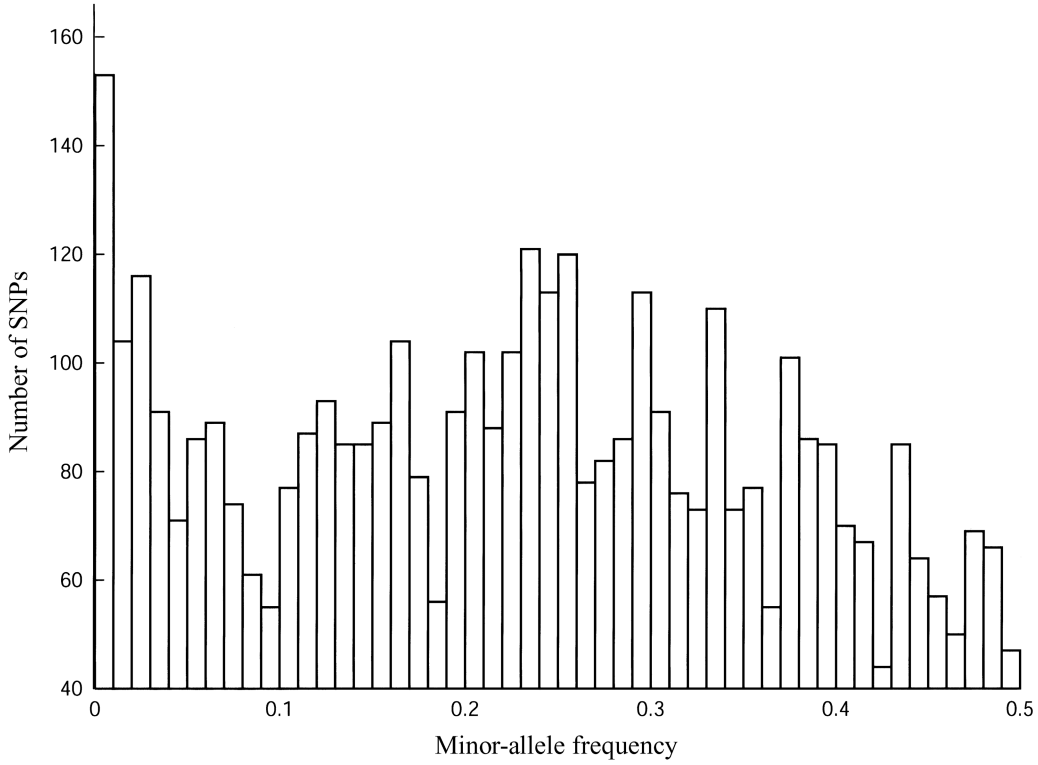

Figure 1 Histogram of minor-allele frequencies for autosomal SNPs. The minor-allele frequency was calculated for each autosomal SNP, and the number of SNPs whose estimated minor-allele frequencies were within an interval is shown.

In the other 174 blocks, neither the 5′ nor the 3′ ends were the same as the edges of the regions. Among these, both ends were unambiguously mapped in 97 blocks, and only those locations were used for estimating the length of the block. The sizes of the blocks were 0.03–137.47 kb, with a mean of 13.54 kb (SD 16.88 kb) and a median of 8.79 kb. The largest block, with a length of 137.47 kb, extended from *OAT2* to *ABCC10,* and the second largest block, with a length of 77.45 kb, was within *ALDH1A2*.

The sizes of the regions between blocks were also calculated when the SNPs were unambiguously mapped. The sizes of the regions between the blocks were 0.06–82,330.88 kb, with a mean of 5,436.68 kb (SD 12,855.24 kb) and a median of 11.83 kb. From these data, we estimated that the blocks make up ~0.24% and ~42.6%, respectively, of the examined sequence when calculated using the median and average physical distances of block and interblock regions.

Number of Major Haplotypes within a Block

One of the benefits of the haplotype blocks is that most of the haplotypes within a block are explained by a limited number of the major haplotypes, even though the total number of possible haplotypes within the block is large. We calculated, for each block, the number of major haplotypes that explain ⩾90% and ⩾95% of all the haplotypes within the block. The results showed that 3.4916 ± 1.0613 (median 3) and 4.2865 ± 1.5239 (median 4) major haplotypes explained ⩾90% and ⩾95%, respectively, of all the haplotypes within a block.

Number of htSNPs Required to Represent Major Haplotypes within a Block

If the number of the major haplotypes that explain most of the haplotypes within a block is limited, then a small number of htSNPs are likely to represent such major haplotypes. We selected htSNPs that represented all of the major haplotypes for each haplotype block. When such htSNPs were used, all of the major haplotypes could be distinguished from each other. This means that additional markers did not substantially increase the percentages of the haplotypes explained by the major haplotypes.

For the major haplotypes that explained ⩾90% of the haplotypes, the number of htSNPs required was 2.3680 ± 0.9403 (median 2), whereas it was 2.9037 ± 1.1440 (median 3) for those that explained ⩾95% of the haplotypes. When htSNPs were selected from the 3,132

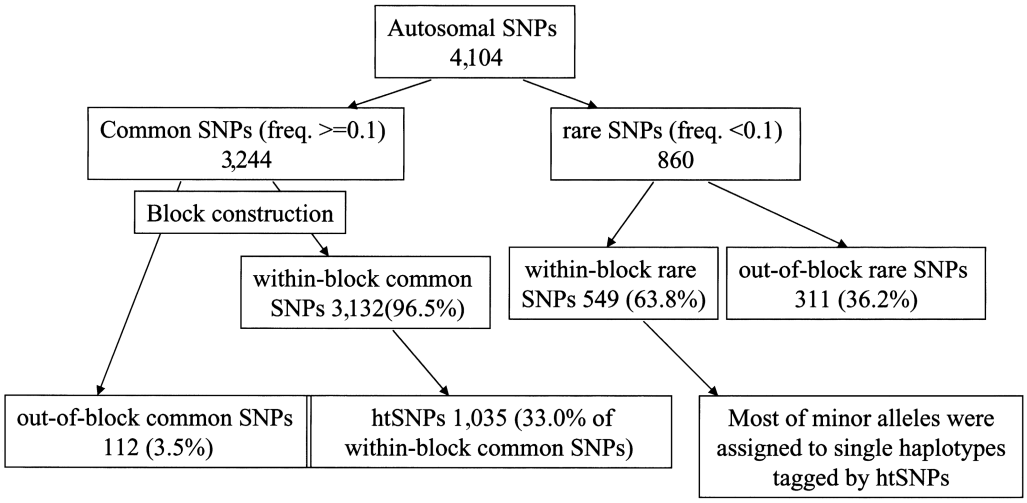

Figure 2 Statistics of 4,104 autosomal SNPs

within-block common SNPs, 1,035 htSNPs were required (fig. 2).

Analysis of the Relationship between Uncommon SNPs and Major Haplotypes

Because the SNPs that we used were discovered using DNA from 48 subjects, the majority of the uncommon SNPs should have been missed. Nonetheless, the population frequencies of some SNPs included in the present study were low, as is shown in figure 1, because sampling is a stochastic process. Note that the frequencies of the minor alleles of even the uncommon SNPs in our data are more accurate than those in previous studies, because our sample size was large. Thus, our data may, for the first time, provide good material with which to examine comprehensively the relationship between uncommon SNPs and major haplotypes.

If we define uncommon SNPs as those with minor-allele frequencies <0.1, 860 SNPs were uncommon (fig. 2). Among these, 549 (63.8%) were within the blocks, and the remaining 311 (36.2%) SNPs were out of the blocks (fig. 2). The assignment of each uncommon SNP to major haplotypes was examined as described in the "Material and Methods" section.

We selected one of the uncommon SNPs (frequency <0.1) from the block, calculated $P(A_i \mid X)$ for all values of i, and selected values of j that gave maximum $P(A_j \mid X)$, as described in the "Material and Methods" section. We selected each of all uncommon SNPs within the block and calculated $P(A_j \mid X)$.

The detailed procedure for calculating $P(A_j \mid X)$ and analyzing the assignment of the minor alleles of an uncommon SNP to htSNP-tagged haplotypes is described in figure 3. In this illustration, a block (block 4) in the *DPYD* (dihydropyrimidine dehydrogenase) gene was used as an example.

$P(A_j \mid X)$, thus calculated, was plotted against the frequency of the uncommon SNP, $P(X)$, for all uncommon SNPs within the blocks. Figure 4 shows the relationship between $P(X)$ and $P(A_j \mid X)$. The results indicated that, irrespective of $P(X)$, $P(A_j \mid X)$ was close to 1 in most cases. Thus, the mean $\pm$ SD for $P(A_j \mid X)$ was 0.943 ± 0.117, and it was >0.9 in 83.9% (459) of the uncommon SNPs. These data indicate that, in most cases, each uncommon SNP is assigned to a single incomplete haplotype defined by the alleles at htSNPs. Then, we calculated $P(X \mid A_j)$ for all uncommon SNPs within the blocks and plotted it against $P(X)$ (fig. 5). The detailed procedure to calculate $P(X \mid A_j)$ is illustrated in figure 3. The results indicated that $P(X \mid A_j)$ was positively correlated with $P(X)$ in the region $(0 < P(X) < 0.03$; $P < .000001$; $n = 233$ [Spearman's rank correlation coefficient]), whereas it was not in the other region $(0.03 \leqslant P(X) < 1$; $P = .050$; $n = 314)$.

In the above experiments, we examined the relationship between uncommon SNPs and incomplete haplotypes through use of the haplotype blocks determined by a method described elsewhere (Zhu et al. 2003). We also used the haplotype blocks determined by the other method (Gabriel et al. 2002) and compared the results. The total number of blocks determined by the method of Gabriel et al. (2002) was 420, which is a little larger than total obtained by using the other method. The length of the block was 11.01 ± 8.41 kb (0.021–166.33 kb; median 5.15 kb). Thus, the lengths of the blocks determined using the method of Gabriel et al. (2002) were, on average, shorter than those determined using the method of Zhu et al. (2003). Even though the hap-

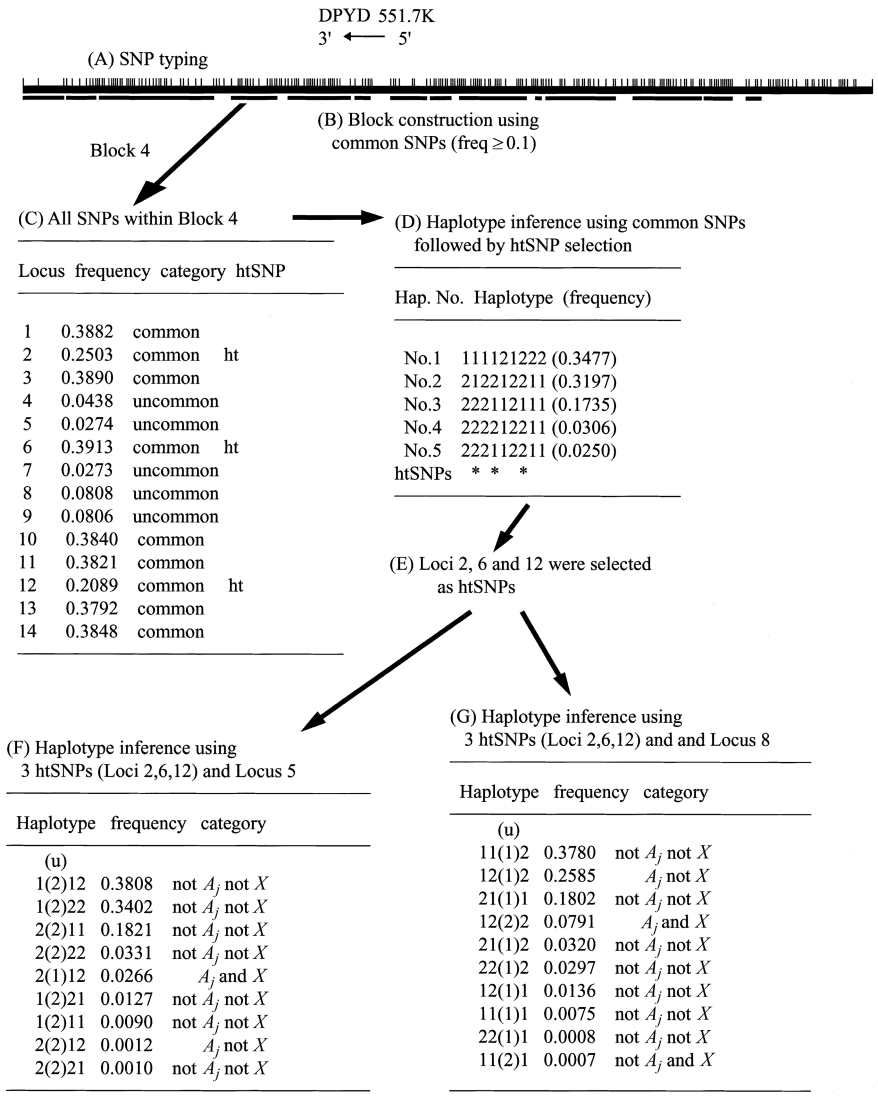

Figure 3 Illustration of the procedure for the analysis of the assignment of uncommon SNPs within an LD block to haplotypes tagged by htSNPs. *A,* Genotype data from the *DPYD* (dihydropyrimidine dehydrogenase) gene on 1p. *B,* Common SNP loci (minor-allele frequencies ≥0.1) chosen from among all SNP loci and used to construct LD blocks, as described in the "Material and Methods" section. *C,* Block 4, with nine common SNPs, was used as an example from among a total of 14 blocks constructed. This block included five uncommon SNPs in addition to the nine common SNPs. The htSNPs ("ht") selected later (*E*) are also indicated. *D,* Haplotypes inferred using only the nine common SNPs, as described in the "Material and Methods" section. A haplotype is expressed as a list of either "1" or "2" alleles, along with the estimated frequency in the parentheses. *E,* htSNPs selected from the inferred haplotype data (*D*), as described in the "Material and Methods" section. For block 4, loci 2, 6, and 12 (indicated by asterisks in panel D) were selected as htSNPs. *F,* One of the uncommon SNPs within block 4 (loci 4, 5, 7, 8, or 9 in panel C) chosen for the analysis of the assignment, in addition to the three htSNPs selected. In this case, locus 5 was chosen as the uncommon SNP. By use of the genotype data from loci 2, 5, 6, and 12, haplotypes were inferred as described in the "Material and Methods"section. An inferred haplotype was described as a list of either one or two alleles at four SNP loci. In the list, the allele at the uncommon SNP (locus 5) is shown in parentheses ("u"). There was only one haplotype with (1), the minor allele of the uncommon SNP—that is, 2(1)12. Therefore, the haplotype 2(−)12 was judged to be A_j, the htSNP-tagged haplotype to which the majority of the minor alleles of an uncommon SNP were assigned. All haplotypes were categorized according to the presence (X) or absence ($\bar{X}$) of (1) and the presence (A_j) or absence ($\bar{A}_j$) of the haplotype 2(−)12. Thus, there are four categories: (*a*) A_j,X (A_j and X), (*b*) $A_j,\bar{X}$ (A_j, not X), (*c*) $\bar{A}_j,X$ (X, not A_j), and (*d*) $\bar{A}_j,\bar{X}$ (not A_j, not X). The frequencies of the haplotypes in the same category were summed to calculate the estimated probabilities $P(A_j,X)$, $P(A_j,\bar{X})$, $P(\bar{A}_j,X)$, and $P(\bar{A}_j,\bar{X})$. For locus 5, the following estimated probabilities were obtained: $P(A_j,X) = 0.0266$, $P(A_j,\bar{X}) = 0.0012$, $P(\bar{A}_j,X) = 0$, and $P(\bar{A}_j,\bar{X}) = 0.9722$. Thus, for locus 5, $P(A_j \mid X) = P(A_j,X)/[P(A_j,X) + P(\bar{A}_j,X)] = 1$, thereby indicating that all the minor alleles at locus 5 were assigned to the haplotype 2(−)12 (A_j). *G,* Locus 8 chosen as the uncommon SNP. By use of the genotype data from loci 2, 6, 8, and 12, haplotypes were inferred as described in the "Material and Methods" section. An inferred haplotype was described as a list of either one or two alleles at four SNP loci, and the allele at the uncommon SNP (locus 8) is shown in parentheses ("u"). There were only two haplotypes with (2), the minor allele of the uncommon SNP. By comparing the frequencies, the haplotype 12(2)2, rather than 11(2)1, was judged to be A_j, the htSNP-tagged haplotype to which the majority of the minor alleles of an uncommon SNP were assigned. All haplotypes were categorized according to the presence (X) or absence ($\bar{X}$) of (2) and the presence (A_j) or absence ($\bar{A}_j$) of the haplotype 12(−)2. The frequencies for the haplotypes in the same category were summed to calculate the estimated probabilities $P(A_j,X)$, $P(A_j,\bar{X})$, $P(\bar{A}_j,X)$, and $P(\bar{A}_j,\bar{X})$. For locus 8, the following estimated probabilities were obtained: $P(A_j,X) = 0.0791$, $P(A_j,\bar{X}) = 0.2585$, $P(\bar{A}_j,X) = 0.0007$, and $P(\bar{A}_j,\bar{X}) = 0.6617$. Thus, for locus 8, $P(A_j \mid X) = P(A_j,X)/[P(A_j,X) + P(\bar{A}_j,X)] = 0.991$, thereby indicating that the majority of the minor alleles at locus 8 were assigned to the haplotype 12(−)2 (A_j).

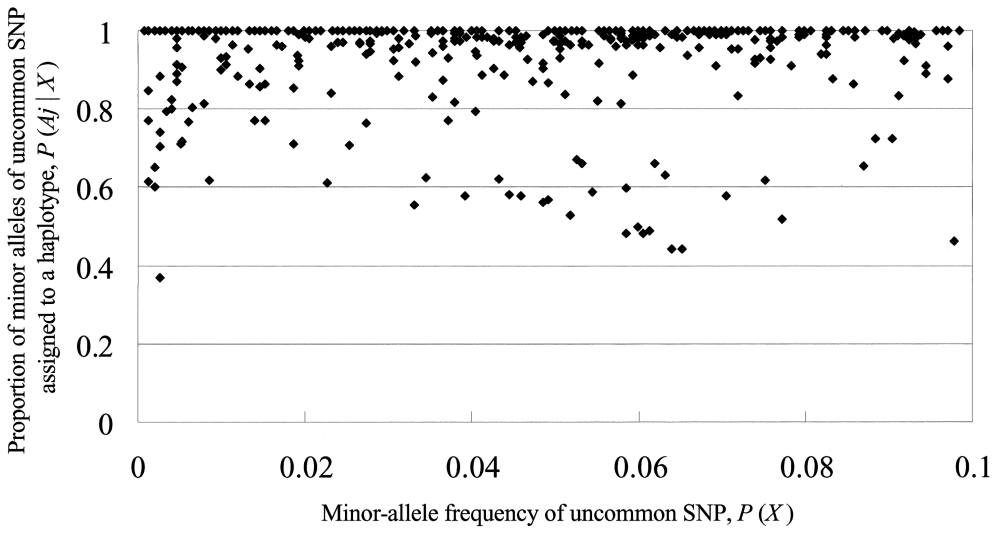

Figure 4 Proportion of minor alleles (X) of an uncommon SNP assigned to a haplotype A_j tagged by htSNPs—that is, $P(A_j \mid X)$. The procedure to calculate $P(A_j \mid X)$ is described in detail in figure 3. $P(X)$ represents the frequency of the minor allele of an uncommon SNP.

lotype blocks determined by the method of Gabriel et al. (2002) were a little different from those determined using the method of Zhu et al. (2003), the proportion of an uncommon SNP assigned to a haplotype—that is, $P(A_j \mid X)$—was also close to 1 (0.954 ± 0.114 for the data from 353 uncommon SNPs within the blocks). When uncommon SNPs located outside of the blocks but adjacent to them were examined for assignment to the incomplete (htSNP-tagged) haplotypes within the blocks, the majority of the minor alleles of uncommon SNPs were not necessarily assigned to single incomplete haplotypes (data not shown). Therefore, the assignment of the majority of the minor alleles to single incomplete haplotypes is a characteristic of uncommon SNPs within the blocks but not of those outside of the blocks.

When X was first generated by a mutation, $P(X \mid A_j)$ was probably very low, and $P(A_j \mid X)$ was probably 1. As time goes on, $P(X)$ and $P(X \mid A_j)$ will increase, and $P(A_j \mid X)$ will decrease if X does not disappear. Therefore, the data for $P(A_j \mid X)$ and for $P(X \mid A_j)$, when $P(X) < 0.03$, reflect the status when X was first generated.

Probability to Detect Phenotype-Associated Uncommon SNPs with Haplotypes Tagged by htSNPs

Haplotypes constructed using only htSNPs (incomplete haplotypes) are expected to be useful in the identification of major haplotypes. However, it is not clear how useful htSNPs and incomplete haplotypes are for the identification of uncommon phenotype-associated SNPs. We studied this problem through use of our haplotype data.

It is known that severe adverse effects from some drugs can occur in homozygous persons who have enzyme deficiencies. In those cases, the causative minor allele X is expected to be elevated in cases—that is, $P(X \mid \psi)$ is expected to be much higher than $P(X \mid \bar{\psi})$, which is low. However, the above condition is not sufficient for incomplete haplotypes to be useful in detecting X. For the association studies using incomplete haplotypes but not uncommon phenotype-associated SNPs to be useful, $P(A_j \mid X)$ should be different from $P(A_j \mid \bar{X})$ (see appendix A).

Therefore, we calculated $P(A_j \mid \bar{X})/P(A_j \mid X)$ for all X within the blocks and plotted the results against $P(X)$ in figure 6. The results indicated that $P(A_j \mid X)$ is much higher than $P(A_j \mid \bar{X})$ in many of the cases, especially when $P(X)$ is >0.02. In fact, the mean $\pm$ SD of $P(A_j \mid \bar{X})/P(A_j \mid X)$ was 0.200 ± 0.230. Therefore, A_j may be used as a marker of X in many cases in association studies searching for X.

The probability of detecting significance through use of A_j instead of X in association studies depends on factors such as $P(X \mid \psi)$, $P(X \mid \bar{\psi})$, $P(A_j \mid \bar{X})/P(A_j \mid X)$, and M_1,M_2, the sizes of affected and control samples, as shown in appendix A. Extensive simulation studies were performed using various parameter sets to address this point, as described in the "Material and Methods" section and in appendix A.

Figure 7 shows an example of the results of the simulation. The graph shows the empirical probability of significance (power) of the test when $r = P(X \mid \psi)/P(X \mid \bar{\psi}) = 8$, $M_1 = 50$, and $M_2 = 500$. The significance level was set at .01. We extensively examined the power of the test, through use of different param-

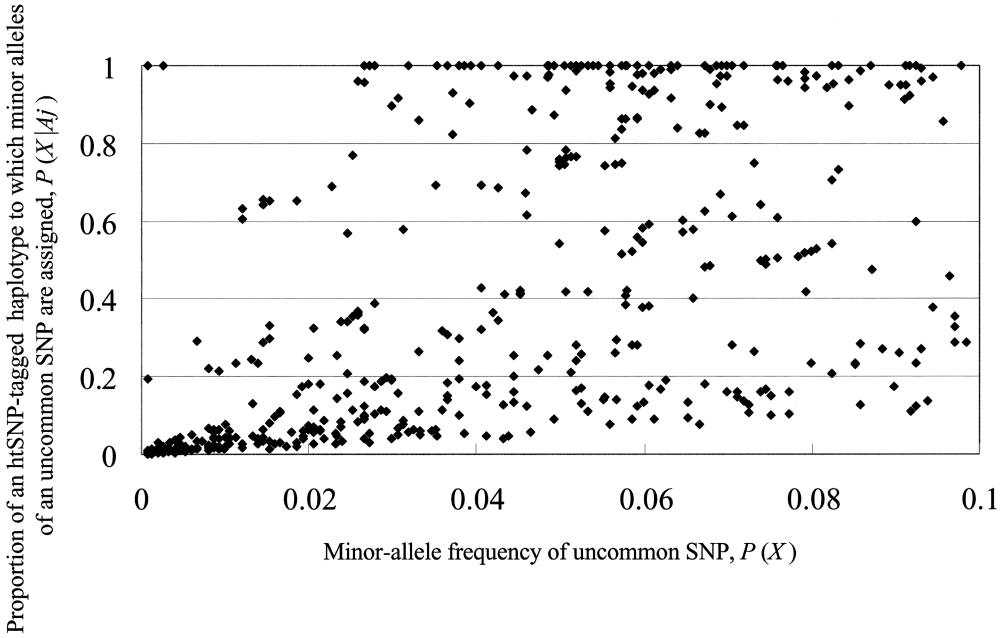

Figure 5 Proportion of htSNP-tagged haplotype A_j containing minor alleles (X) of uncommon SNPs—that is, $P(X \mid A_j)$. $P(X \mid A_j)$ was calculated using $P(X \mid A_j) = P(A_j,X)/[P(A_j,X) + P(A_j,\bar{X})]$. The procedure to calculate $P(A_j,X)$ and $P(A_j,\bar{X})$ is described in detail in figure 3.

eter sets (r = 0.2, 0.5, 3, 5, 8, 10, and 15; M_1 = 50, 100, and 300; and M_2 = 50, 100, and 500). Although both the models and the parameter sets that we examined are still limited, our results suggest that the phenotype-associated X is likely to be detected by A_j when $P(X)$, r, M_1, and M_2 are sufficiently large. For example, under the conditions used in figure 7, the mean ± SD power was 0.239 ± 0.313 (n = 233) when $P(X) < 0.03$, but it was 0.885 ± 0.251 (n = 314) when $P(X) \geq 0.03$. The power depended on $P(X)$ and on such parameters as r, M_1, and M_2, and it increased when either of those parameters or $P(X)$ (the frequency of the minor allele of the uncommon SNP) was increased. Under the conditions used in figure 7, most of the uncommon SNPs within the blocks can be identified by htSNP-tagged haplotypes when the frequencies of the uncommon SNPs are >0.03.

Discussion

In the present investigation, we estimated the frequencies of SNPs, constructed haplotype blocks, estimated the frequencies of haplotypes within the blocks, and selected htSNPs representing the major haplotypes, through use of genotype data at 4,190 SNPs in 199 drug-related genes from 752 subjects. Thereafter, we analyzed the association between uncommon SNPs within the blocks and the haplotypes constructed with htSNPs. The majority of the minor alleles of the uncommon SNPs were assigned to single major haplotypes when the uncommon SNPs were within blocks. The results of simulation studies suggested that haplotype analysis that uses htSNPs may be useful in the detection of uncommon phenotype-associated SNPs if (1) the frequencies of the SNPs are higher in affected populations than in control populations, (2) the SNPs are within the blocks, and (3) the frequencies of the SNPs are >0.03.

As stated in the "Material and Methods" section, the relationship among a haplotype H_i (a complete haplotype), a haplotype constructed with htSNPs A_i (an incomplete haplotype), and the minor allele of an uncommon SNP X within a block can be described by defining the above concepts as events on the sample space Ω, the set of all complete haplotypes in a population. A_j is defined as the incomplete haplotype for which $P(A_j \mid X)$ is the maximum among all A_i for the block. Various conditional probabilities were calculated from the joint probabilities $P(X,A_j)$, $P(X,\bar{A}_j)$, $P(\bar{X},A_j)$, and $P(\bar{X},\bar{A}_j)$. These joint probabilities were estimated by the haplotype-inference algorithm, using genotype data for all of the htSNPs and an uncommon SNP within a block.

The results of our analysis indicated that $P(A_j \mid X)$ was close to 1 in most of the cases, irrespective of $P(X)$, the minor-allele frequency of the uncommon SNP (fig. 4). When X was first generated by a mutation, $P(A_j \mid X)$ was probably 1. As time goes on, $P(A_j \mid X)$ is likely to decrease because of recombination within the blocks. Our results

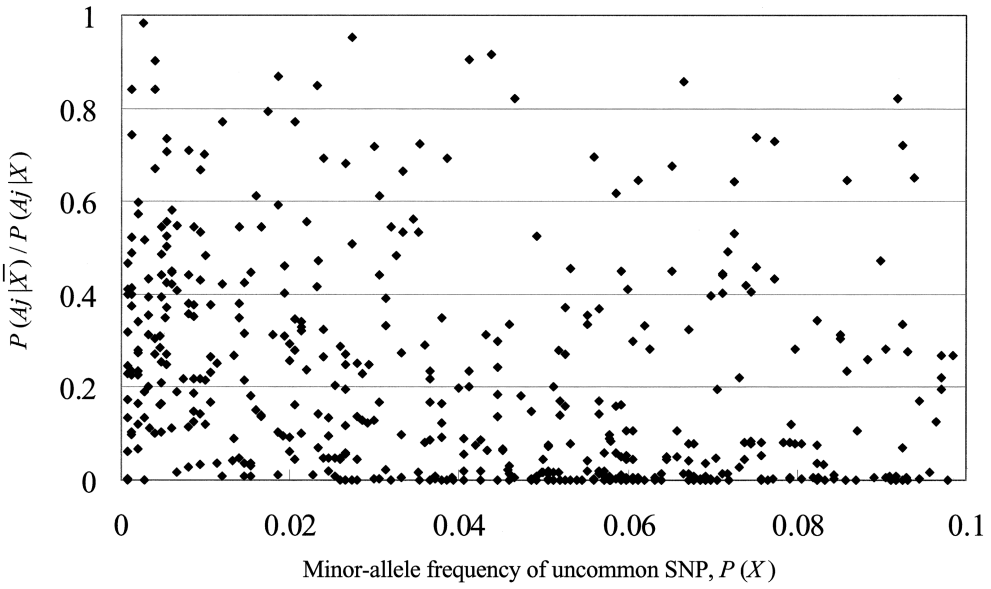

Figure 6 Ratio of the proportion of a haplotype tagged by htSNPs (A_j) containing the major alleles ($\bar{X}$) of an uncommon SNP to the proportion of the same haplotype containing the minor alleles (X)—that is, the ratio $P(A_j \mid \bar{X})/P(A_j \mid X)$.

support the hypothesis that only infrequent recombinations occur within haplotype blocks and that out-of-block regions, but not block regions, include recombinational "hotspots."

Information about SNPs and their frequency in drug-related genes will be very important for population-based pharmacogenetic studies, as well as for the development of personalized drug therapy. Our results, which are summarized as the statistics for 4,104 autosomal SNPs in figure 2, can suggest a method for the optimization of population-based pharmacogenetic studies. Thus, from 3,244 common SNPs, we could select 1,035 htSNPs (31.9%) that represent most of the major haplotypes within the blocks. Because ~64.2% of uncommon SNPs were within such blocks and the majority of their minor alleles were assigned to single incomplete haplotypes defined by htSNPs, these 1,035 htSNPs can be used to search for both common and uncommon SNPs associated with phenotypes if the phenotype-related SNPs are within the blocks. Only 112 (3.5%) of 3,244 common SNPs were located outside of the blocks (fig. 2), and we can use all of these to search for phenotype-related common SNPs located outside of the blocks. Therefore, for optimal population-based pharmacogenetic studies, genotyping a total of 1,147 (1,035 + 112) common SNPs is required, which represents ~35.2% (1,144/3,244) of all of the common SNPs found (fig. 2).

It is obvious that the most difficult problem is uncommon SNPs that are phenotype-related and located outside of the blocks. Some of these may be in weak LD with adjacent out-of-block common SNPs and may be detected using out-of-block common SNPs as tags. However, the power of this method should not be high.

An international project to construct a genomewide haplotype map has been started (the HapMap project) (Couzin 2002); however, it is focused on common SNPs. The common SNPs are, of course, very important in drug reactions. However, uncommon SNPs are also likely to be important, especially for uncommon severe adverse reactions. For example, if an uncommon nonsynonymous SNP is responsible for a severe adverse reaction, then the frequency of a major haplotype to which most of the minor alleles of the uncommon SNP are assigned may have increased among the affected subjects. Such a haplotype may be identified using htSNPs.

Thus, we have provided a database of SNPs and haplotype blocks of 199 drug-related genes using samples from 752 subjects. During the analysis of the data, we noticed, using the programs for haplotype inference, that most of the minor alleles of each uncommon SNP within a block are assigned to a single haplotype. These data suggest that the construction of haplotype blocks and the selection of htSNPs may be useful in identifying not only common SNPs but also unidentified uncommon SNPs that are associated with phenotypes such as diseases and drug reactions. Although the results of previous studies suggested that EM-based haplotype inference is accurate (Long et al. 1995; Kitamura et al. 2002), our data are worth being reexamined on the basis of data from molecular haplotyping. In addition, although

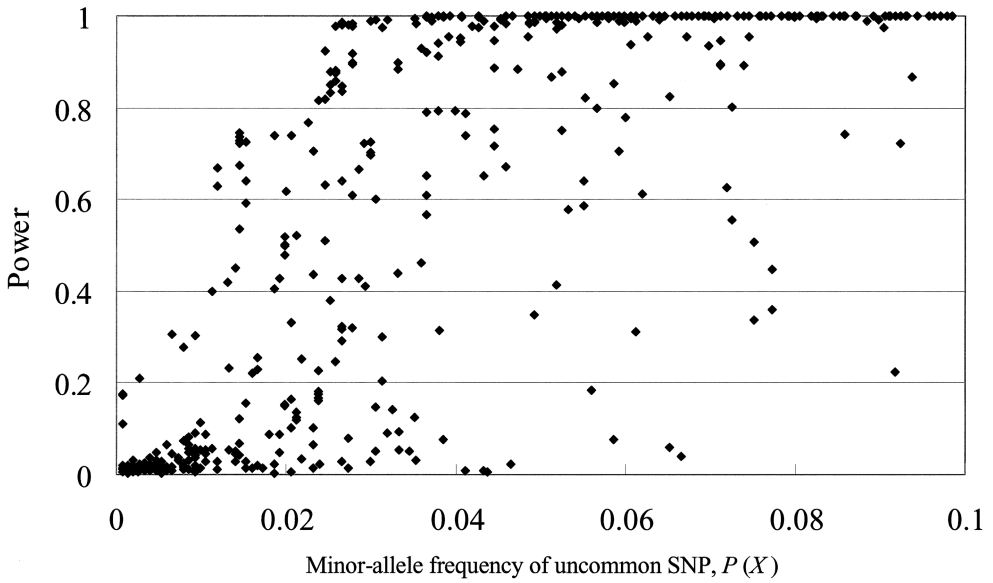

Figure 7 Probability of the significance for the tests comparing the frequencies of the haplotypes tagged by htSNPs between affected and control populations when the phenotype-associated uncommon SNP differs in frequency between the two populations. The ratio $r = P(X \mid \psi)/P(X \mid \bar{\psi})$ was set at 8, and the significance level was set at .01. The numbers of the subjects in affected and control samples (M_1 and M_2) were set at 50 and 500, respectively. The detailed procedure for the simulation is described in appendix A.

we used, as have many other researchers, the threshold of 0.1 to differentiate common from uncommon SNPs throughout the present article, the validity of this threshold should be examined in future studies. Ke et al. (2004) reported that the pattern of LD changed when the density of SNPs was varied. When they used denser SNP maps, overall sequence coverage in LD blocks and block boundaries varied substantially (Ke et al. 2004). We are now extensively studying the effects of changing thresholds for differentiating common from uncommon SNPs on haplotype blocks. When the threshold becomes lower, the SNP maps become denser. Preliminary results suggested that a lower threshold leads to the increase in the overall sequence coverage in haplotype blocks. The effects of the lower threshold on the lengths of haplotype blocks seemed complicated. Some blocks became longer because of the addition of new common SNPs to the boundaries of the blocks, whereas others became shorter, since the introduction of new common SNPs resulted in the breakage of the blocks. However, the assignment of the majority of the minor alleles of uncommon SNPs within blocks to single haplotypes was observed even when the thresholds for differentiating common from uncommon SNPs were changed from 0.05 to 0.15. Further studies are necessary to determine the precise effects of the change in the threshold for differentiating the common from uncommon SNPs.

Acknowledgments

The present study was supported by grants from the Pharma SNP Consortium and the Research for the Future Program from the Japan Society for the Promotion of Science.

Appendix A

In the simulation, the locus involving X, a set of minor alleles of an uncommon SNP within an LD block, was assumed to be the only locus directly associated with the phenotype. Even though A_j, the htSNP-tagged haplotype to which the majority of the members of X are assigned, may be associated with ψ (the set of complete haplotypes in affected subjects), the association is not direct but instead exists through an association between X and the phenotype. Because X is associated with the phenotype, the frequencies are expected to be different between affected and unaffected subjects—that is, $P(X \mid \psi) \neq P(X \mid \bar{\psi})$, where $\bar{\psi}$ denotes the complement of ψ. In a case-control study, the difference between those frequencies is tested. Let r denote the ratio $r = P(X \mid \psi)/P(X \mid \bar{\psi})$. We tested whether A_j could be used as an effective marker to detect X. Because the frequencies of X are different between ψ and $\bar{\psi}$

and A_j is tightly associated with X, the frequencies of A_j are likely to be different between ψ and $\bar{\psi}$. The frequencies of A_j are obtained as follows:

$$P(A_j \mid \psi) = P(X \mid \psi)P(A_j \mid X,\psi) + P(\bar{X} \mid \psi)P(A_j \mid \bar{X},\psi) \tag{A1}$$

and

$$P(A_j \mid \bar{\psi}) = P(X \mid \bar{\psi})P(A_j \mid X,\bar{\psi}) + P(\bar{X} \mid \bar{\psi})P(A_j \mid \bar{X},\bar{\psi}) \ . \tag{A2}$$

Because the locus with X is the only locus directly associated with the phenotype, and because the association between A_j and the phenotype is only through the former association, the following equations hold:

$$P(\psi \mid X,A_j) = P(\psi \mid X,\bar{A}_j) = P(\psi \mid X) \tag{A3}$$

and

$$P(\bar{\psi} \mid X,A_j) = P(\bar{\psi} \mid X,\bar{A}_j) = P(\bar{\psi} \mid X) \ .$$

When Bayes's theorem is used, the following equation is obtained:

$$P(A_j \mid X,\psi) = \frac{P(A_j \mid X)P(\psi \mid X,A_j)}{P(A_j \mid X)P(\psi \mid X,A_j) + P(\bar{A}_j \mid X)P(\psi \mid X,\bar{A}_j)} \ . \tag{A4}$$

From equation (A3), equation (A4) becomes

$$P(A_j \mid X,\psi) = \frac{P(A_j \mid X)P(\psi \mid X)}{P(A_j \mid X)P(\psi \mid X) + P(\bar{A}_j \mid X)P(\psi \mid X)} = P(A_j \mid X) \ .$$

Similarly, we obtain $P(A_j \mid X,\bar{\psi}) = P(A_j \mid X)$, and $P(A_j \mid \bar{X},\psi) = P(A_j \mid \bar{X},\bar{\psi}) = P(A_j \mid \bar{X})$. Thus, A_j is independent of ψ or $\bar{\psi}$ conditional on X or $\bar{X}$. This is important, because we calculate $P(A_j \mid \psi)$ and $P(A_j \mid \bar{\psi})$ from equations (A1) and (A2).

For the simulation in the present study, $P(X)$ was used in place of $P(X \mid \bar{\psi})$, because we consider the case where $P(\psi)$ is small and $P(X) \simeq P(X \mid \bar{\psi})$. Thus, the frequency of X in the control group was assumed to be the same as the population frequency. Therefore, equations (A1) and (A2) become

$$P(A_j \mid \psi) = rP(X)P(A_j \mid X) + [1 - rP(X)]P(A_j \mid \bar{X}) \tag{A5}$$

and

$$P(A_j \mid \bar{\psi}) = P(X)P(A_j \mid X) + [1 - P(X)]P(A_j \mid \bar{X}) \ . \tag{A6}$$

$P(A_j \mid X)$ and $P(A_j \mid \bar{X})$ for each uncommon SNP were calculated as $P(A_j \mid X) = P(A_j,X)/[P(A_j,X) + P(\bar{A}_j,X)]$, and $P(A_j \mid \bar{X}) = P(A_j,\bar{X})/[P(A_j,\bar{X}) + P(\bar{A}_j,\bar{X})]$. The procedure to calculate $P(A_j,X)$,$P(A_j,\bar{X})$,$P(\bar{A}_j,X)$ and $P(\bar{A}_j,\bar{X})$ for each uncommon SNP within an LD block was as described in the "Material and Methods" section.

Before the simulation, a set of values were given to the ratio r and the numbers of the subjects in the sample, M_1 and M_2. For the simulation, $2M_1$ haplotypes were generated for the affected subjects from a binary distribution with a frequency parameter of $P(A_j \mid \psi)$, and $2M_2$ haplotypes were generated for the control subjects from a binary distribution with a frequency parameter of $P(A_j \mid \bar{\psi})$. Pearson's χ^2 test was used to detect the difference between $P(A_j \mid \psi)$ and $P(A_j \mid \bar{\psi})$, using the 2×2 contingency table thus obtained. The proportion of the iterations with $P < .01$ was interpreted as the empirical power. A total of 5,000 iterations was performed for an uncommon SNP.

This test aims to detect the difference between $P(A_j \mid \psi)$ and $P(A_j \mid \bar{\psi})$. Therefore, the ratio $P(A_j \mid \psi)/P(A_j \mid \bar{\psi})$ plays a critical role. From equations (A5) and (A6), this ratio is

$$P(A_j \mid \psi)/P(A_j \mid \bar{\psi}) = \frac{\left\{r + \left[\frac{1}{P(X)} - r\right]\frac{P(A_j \mid \bar{X})}{P(A_j \mid X)}\right\}}{\left\{1 + \left[\frac{1}{P(X)} - 1\right]\frac{P(A_j \mid \bar{X})}{P(A_j \mid X)}\right\}} . \quad \text{(A7)}$$

Therefore, the ratio (A7) depends on $P(A_j \mid \bar{X})/P(A_j \mid X)$. When $P(A_j \mid \bar{X})/P(A_j \mid X) = 1$, the ratio (A7) becomes 1, and one cannot detect the difference between $P(A_j \mid \psi)$ and $P(A_j \mid \bar{\psi})$. The software for the simulation, ANASSIGN, written in C language, was written by the authors and will be provided on request.

Electronic-Database Information

The URL for data presented herein is as follows:

Genstat, http://genstat.net/haplotypeblock/listofgenes.html (for a list of genes included in the present study)

References

Avi-Itzhak HI, Su X, De La Vega FM (2003) Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. Pac Symp Biocomput 2003:466–477

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173–15177

Couzin J (2002) Human genome: HapMap launched with pledges of $100 million. Science 298:941–942

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

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 USA 97:10483–10488

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

Evans WE, Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286:487–491

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

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

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 26:163–175

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Iida A, Saito S, Sekine A, Mishima C, Kitamura Y, Kondo K, Harigae S, Osawa S, Nakamura Y (2002*a*) Catalog of 605 single-nucleotide polymorphisms (SNPs) among 13 genes encoding human ATP-binding cassette transporters:ABCA4, ABCA7, ABCA8, ABCD1, ABCD3, ABCD4, ABCE1, ABCF1, ABCG1, ABCG2, ABCG4, ABCG5, and ABCG8. J Hum Genet 47:285–310

——— (2002*b*) Catalog of 86 single-nucleotide polymorphisms (SNPs) in three uridine diphosphate glycosyltransferase genes: UGT2A1, UGT2B15, and UGT8. J Hum Genet 47:505–510

——— (2003) Catalog of 668 SNPs detected among 31 genes encoding potential drug targets on the cell surface. J Hum Genet 48:23–46

Iida A, Saito S, Sekine A, Mishima C, Kondo K, Kitamura Y, Harigae S, Osawa S, Nakamura Y (2001*a*) Catalog of 258 single-nucleotide polymorphisms (SNPs) in genes encoding three organic anion transporters, three organic anion-transporting polypeptides, and three NADH:ubiquinone oxidoreductase flavoproteins. J Hum Genet 46:668–683

Iida A, Sekine A, Saito S, Kitamura Y, Kitamoto T, Osawa S, Mishima C, Nakamura Y (2001*b*) Catalog of 320 single nucleotide polymorphisms (SNPs) in 20 quinone oxido-

reductase and sulfotransferase genes. J Hum Genet 46:225–240

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

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

Jurinke C, van den Boom D, Cantor CR, Koster H (2002) The use of MassARRAY technology for high throughput genotyping. Adv Biochem Eng Biotechnol 77:57–74

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

Kitamura Y, Moriguchi M, Kaneko H, Morisaki H, Morisaki T, Toyama K, Kamatani N (2002) Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm. Ann Hum Genet 66:183–193

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

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234–236

Kwiatkowski RW, Lyamichev V, de Arruda M, Neri B (1999) Clinical, genetic, and pharmacogenetic applications of the Invader assay. Mol Diagn 4:353–364

Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 279:1200–1205

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

Meyer UA (2000) Pharmacogenetics and adverse drug reactions. Lancet 356:1667–1671

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603–606

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46:471–477

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. Science 294:1719–1723

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

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

Saito S, Iida A, Sekine A, Ogawa C, Kawauchi S, Higuchi S, Nakamura Y (2002) Catalog of 238 variations among six human genes encoding solute carriers (hSLCs) in the Japanese population. J Hum Genet 47:576–584

Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000) dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28:352–355

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

Tanaka E, Taniguchi A, Urano W, Nakajima H, Matsuda Y, Kitamura Y, Saito M, Yamanaka H, Saito T, Kamatani N (2002) Adverse effects of sulfasalazine in patients with rheumatoid arthritis are associated with diplotype configuration at the N-acetyltransferase 2 gene. J Rheumatol 29:2492–2499

Urano W, Taniguchi A, Yamanaka H, Tanaka E, Nakajima H, Matsuda Y, Akama H, Kitamura Y, Kamatani N (2002) Polymorphisms in the methylenetetrahydrofolate reductase gene were associated with both the efficacy and the toxicity of methotrexate used for the treatment of rheumatoid arthritis, as evidenced by single locus and haplotype analyses. Pharmacogenetics 12:183–190

Zhu X, Yan D, Cooper RS, Luke A, Ikeda MA, Chang YP, Weder A, Chakravarti A (2003) Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. Genome Res 13:173–181