Acknowledgments

We thank all the family members for participating in this study. We also thank the Commission for Families and Children of Orange County for its support of our clinical work. We appreciate Dr. Takeyori Saheki for useful comments on metabolic decompensation. This study was performed at the Advanced Medical Research Center, Yokohama City University.

Disclosure statement: The authors have no conflict of interest to declare.

References

Barel O, Shorer Z, Flusser H, Ofir R, Narkis G, Finer G, Shalev H, Nasasra A, Saada A, Birk OS. 2008. Mitochondrial complex III deficiency associated with a homozygous mutation in UQCRQ. Am J Hum Genet 82:1211–1216.

Benit P, Lebon S, Rustin P. 2009. Respiratory-chain diseases related to complex III deficiency. Biochim Biophys Acta 1793:181–185.

de Lonlay P, Valnot I, Barrientos A, Gorbatyuk M, Tzagoloff A, Taanman JW, Benayoun E, Chretien D, Kadhom N, Lombes A, de Baulny HO, Niaudet P, et al. 2001. A mutant mitochondrial respiratory chain assembly protein causes complex III deficiency in patients with tubulopathy, encephalopathy and liver failure. Nat Genet 29:57–60.

DiMauro S, Schon EA. 2003. Mitochondrial respiratory-chain diseases. N Engl J Med 348:2656–2668.

Fernandez-Vizarra E, Bugiani M, Goffrini P, Carrara F, Farina L, Procopio E, Donati A, Uziel G, Ferrero I, Zeviani M. 2007. Impaired complex III assembly associated with *BCS1L* gene mutations in isolated mitochondrial encephalopathy. Hum Mol Genet 16:1241–1252.

Ghezzi D, Arzuffi P, Zordan M, Da Re C, Lamperti C, Benna C, D'Adamo P, Diodato D, Costa R, Mariotti C, Uziel G, Smiderle C, et al. 2011. Mutations in TTC*19* cause mitochondrial complex III deficiency and neurological impairment in humans and flies. Nat Genet 43:259–263.

Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A. 2005. Allegro version 2. Nat Genet 37:1015–1016.

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320:369–387.

Haut S, Brivet M, Touati G, Rustin P, Lebon S, Garcia-Cazorla A, Saudubray JM, Boutron A, Legrand A, Slama A. 2003. A deletion in the human QP-C gene causes a complex III deficiency resulting in hypoglycaemia and lactic acidosis. Hum Genet 113:118–122.

Hinson JT, Fantin VR, Schonberger J, Breivik N, Siem G, McDonough B, Sharma P, Keogh I, Godinho R, Santos F, Esparza A, Nicolau Y, et al. 2007. Missense mutations in the BCS*1L* gene as a cause of the Bjornstad syndrome. N Engl J Med 356:809–819.

Iwata S, Lee JW, Okada K, Lee JK, Iwata M, Rasmussen B, Link TA, Ramaswamy S, Jap BK. 1998. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. Science 281:64–71.

Khan S, Vihinen M. 2010. Performance of protein stability predictors. Hum Mutat 31:675–684.

Mitsuhashi S, Hatakeyama H, Karahashi M, Koumura T, Nonaka I, Hayashi YK, Noguchi S, Sher RB, Nakagawa Y, Manfredi G, Goto Y, Cox GA, Nishino I. 2011. Muscle choline kinase beta defect causes mitochondrial dysfunction and increased mitophagy. Hum Mol Genet 20:3841–3851.

Schagger H, Pfeiffer K. 2000. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. EMBO J 19:1777–1783.

Trounce IA, Kim YL, Jun AS, Wallace DC. 1996. Assessment of mitochondrial oxidative phosphorylation in patient muscle biopsies, lymphoblasts, and transmitochondrial cell lines. Methods Enzymol 264:484–509.

Tsurusaki Y, Osaka H, Hamanoue H, Shimbo H, Tsuji M, Doi H, Saitsu H, Matsumoto N, Miyake N. 2011. Rapid detection of a mutation causing X-linked leucoencephalopathy by exome sequencing. J Med Genet 48:606–609.

Visapaa I, Fellman V, Vesa J, Dasvarma A, Hutton JL, Kumar V, Payne GS, Makarow M, Van Coster R, Taylor RW, Turnbull DM, Suomalainen A, et al. 2002. GRACILE syndrome, a lethal metabolic disorder with iron overload, is caused by a point mutation in BCS*1L*. Am J Hum Genet 71:863–876.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 1, 2012
© Mary Ann Liebert, Inc.
Pp. 55–67
DOI: 10.1089/cmb.2010.0227

Population Model–Based Inter-Diplotype Similarity Measure for Accurate Diplotype Clustering

RITSUKO ONUKI,[1] RYO YAMADA,[2] RUI YAMAGUCHI,[3]
MINORU KANEHISA,[1] and TETSUO SHIBUYA[3]

ABSTRACT

Classification of the individuals' genotype data is important in various kinds of biomedical research. There are many sophisticated clustering algorithms, but most of them require some appropriate similarity measure between objects to be clustered. Hence, accurate inter-diplotype similarity measures are always required for classification of diplotypes. In this article, we propose a new accurate inter-diplotype similarity measure that we call the population model-based distance (PMD), so that we can cluster individuals with diplotype SNPs data (i.e., unphased-diplotypes) with higher accuracies. For unphased-diplotypes, the allele sharing distance (ASD) has been the standard to measure the genetic distance between the diplotypes of individuals. To achieve higher clustering accuracies, our new measure PMD makes good use of a given appropriate population model which has never been utilized in the ASD. As the population model, we propose to use an hidden Markov model (HMM)–based model. We call the PMD based on the model the HHD (HIT HMM–based Distance). We demonstrate the impact of the HHD on the diplotype classification through comprehensive large-scale experiments over the genome-wide 8930 data sets derived from the HapMap SNPs database. The experiments revealed that the HHD enables significantly more accurate clustering than the ASD.

Key words: algorithms, statistics, strings, suffix trees.

1. INTRODUCTION

SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) are the most fundamental genetic polymorphisms in human genomes (Kim and Misra, 2007), and classification of individuals with the individual SNPs data is very useful in various kinds of biomedical research, especially in population genetics and genetic epidemiology (Conrad et al., 2006; Jakobsson et al., 2008). Accurate classification of individual SNPs data will help study of genotype variations, especially when different genotypes prevail in different populations or subgroups.

There are various sophisticated clustering methods for general data (not limited for clustering SNPs data), many of which (e.g., Ward's method [Team RDC, 2007; Ward, 1963; Ward and Hook, 1963],

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto Japan.
[2] Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
[3] Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan.

k-Medoid [Kaufman and Rousseuw, 1990], DBSCAN [Ester et al., 1996], and most of the phylogenetic clustering algorithms such as the famous neighbor joining method [Saitou and Nei, 1987]) require appropriate similarity measures between target objects. Designing accurate similarity measure for the objects to be clustered is essential for these similarity-based clustering algorithms.

For SNPs data, there have been proposed various clustering algorithms for clustering haplotypes (i.e., haplotype-alleles, not diplotypes),[1] and various types of similarity measures have been proposed for haplotype data (Jin et al., 2010; Li and Jiang, 2005; Li et al., 2006).[2] But the human genome is diallelic, and in many cases we observe only the unordered (i.e., unphased) pair of alleles at each locus, instead of ordered (i.e., phased) allele data, due to the high costs required for deciphering unphased allele data to accurate phased ones. In this article, we call a phased pair of haplotypes a ''haplotype-diplotype,'' and we call an unphased pair of haplotypes a ''unphased-diplotype.''

Much work has been done on clustering the unphased-diplotype data. They can be categorized into two types: distance-based methods (Bowcock et al., 1994; Gao and Starmer, 2007) and statistics-based methods (Falush et al., 2003; Pritchard et al., 2000). The distance-based methods utilize a distance measure between two objects, while statistics-based methods are based on the statistical behavior of objects. In this article, we focus on the distance-based clustering methods for unphased-diplotype data. Most previous distance-based methods utilize a similarity measure called the allele sharing distance (ASD) (Gao and Martin, 2009; Jakobsson et al., 2008; Mao et al., 2007; Witherspoon et al., 2007) (see Section 2.1.1). The ASD is a simple and straightforward extension of the Hamming distance, and is the most standard and frequently used similarity measure between a pair of unphased-diplotypes.

In genetic analysis, it is very important to consider properties of populations that are different among genetically distinct populations (Beaty et al., 2005; Fallin et al., 2001; Witherspoon et al., 2007). It should also be true with designing similarity measures for unphased-diplotypes. But the measure ASD does not utilize any population information in obtaining the similarity values. Thus, in this article, we will first propose a new similarity measure called the population model-based distance (PMD) for unphased-diplotypes, which incorporates the population information from an appropriate population model. As the model, we will propose to use an hidden Markov model (HMM)–based model predicted by a standard HMM-based phasing software called HIT (Rastas et al., 2005). We call the PMD based on the model the HHD (the HIT HMM-based distance). We will show the superiority of our new measure HHD over the previous standard ASD through comprehensive experiments over the genome-wide HapMap data (International HapMap Consortium, 2005).

The organization of this article is as follows. In Section 2, we describe previous work on which our method is based. In Section 3, we describe our new measure. In Section 4, we compare the ASD and the HHD through comprehensive experiments over large-scale HapMap data sets to evaluate the impact of the HHD. In Section 5, we conclude.

1.1. Notations and definitions

We assume all SNPs are diallelic. We consider n diplotypes over m SNP loci from the same chromosome. These loci are numbered $1, 2, \cdots, m$ in the physical order. A SNP-allele for a SNP locus is an element in set $\mathcal{S} = \{1, 0\}$ where 1 and 0 denote the major and minor SNP-alleles, respectively. A haplotype-allele is a sequence of SNP-alleles and is represented by a sequence in $\mathcal{S}^m$ (e.g., $10101 \in \mathcal{S}^5$). A SNP-diplotype for a SNP locus is an unordered pair of SNP-allele in $\mathcal{D} = \mathcal{S} \times \mathcal{S}$(e.g., $\{0, 1\} \in \mathcal{D}$). An unphased-diplotype is a sequence of SNP-diplotype and is represented by a sequence in $\mathcal{D}^m$(e.g., $\{1, 0\} - \{0, 0\} - \{1, 0\} - \{1, 1\} - \{1, 0\} \in \mathcal{D}^5$). Given unphased-diplotypes, the phasing problem is to find the most probable corresponding haplotype-allele pairs that could have generated the unphased-diplotypes. A phased haplotype-allele pair is called a haplotype-diplotype (e.g., {10010, 00111}).

[1] There are also many algorithms proposed for clustering SNP loci (Yang and Tabus, 2007), instead of individuals, but we do not deal with these problems in this article.

[2] Various inter-population distances have also been proposed (Cornuet et al., 1999), but we will not deal with these in this article.

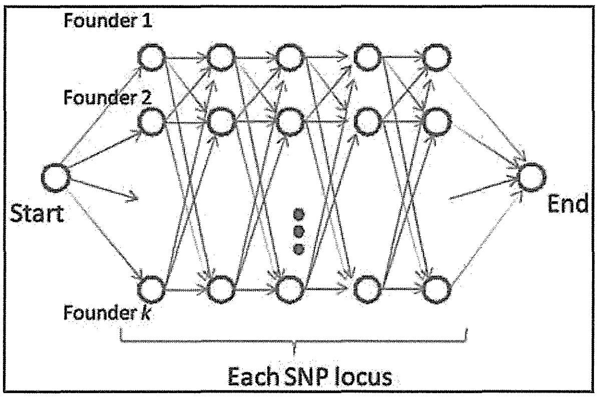

FIG. 1. The HMM model of the HIT. In the HMM, a set of nodes in a row corresponds to states of one founder (i.e., ancestor) haplotype-allele. A set of nodes in a column corresponds to states of one locus. Each node (except for the start and end nodes) emits 1 or 0 with some estimated probabilities, which correspond to the major and minor alleles respectively. A path from the start node to the end node corresponds to a haplotype-allele. The HMM emits a haplotype-diplotype as an unordered pair of two paths from the start node to the end node, randomly based on the probabilities estimated for edges. The observers can only see the unphased-diplotype that corresponds to the emitted haplotype-diplotype.

2. PREVIOUS WORK

In this section, we describe previous work on which our work is based. In Section 2.1, we describe the definitions of measures in previous work (e.g., the ASD). In Section 2.2, we describe the HIT algorithm on which our new distance measure is based. In Section 2.3, we describe a clustering algorithm and an evaluation method for clustering that we will use in the experiments in Section 4.

2.1. *Previous measures for inter-individual genetic distances*

2.1.1. Allele sharing distance. The most standard inter-diplotype distance is the ASD (Gao and Martin, 2009; Jakobsson et al., 2008; Mao et al., 2007; Witherspoon et al., 2007), defined as follows. For two unphased-diplotypes $\mathbf{g}, \mathbf{g}' \in \mathcal{D}^m$ (i.e., m is the number of SNP loci), the ASD between the diplotypes $\mathbf{g}$ and $\mathbf{g}'$ is defined as follows:

$$D(\mathbf{g}, \mathbf{g}') = \frac{1}{2m}\sum_{\ell=1}^{m} d(\mathbf{g}[\ell], \mathbf{g}'[\ell]), \tag{1}$$

where $\mathbf{g}[\ell]$ denotes the ℓ-th SNP-diplotype of unphased-diplotype $\mathbf{g}$, and $d(\mathbf{g}[\ell],\mathbf{g}'[\ell])$ is the number of SNP-alleles which are not shared between $\mathbf{g}$ and $\mathbf{g}'$ at the ℓ-th locus.

2.1.2. Haplotype similarity measure. The most common and simplest measurement for the similarity between DNA sequences, including the haplotype-allele data, is the hamming distance (Cover and Thomas, 1991; Isaev, 2004; Lesk, 2005; Li and Jiang, 2005; Tzeng et al., 2003). For a haplotype-allele $\mathbf{h} \in \mathcal{S}^m$ (where m is the length of $\mathbf{h}$), let $\mathbf{h}[k]$ denote the SNP-allele at the k-th locus of $\mathbf{h}$. The hamming distance between two haplotype-alleles $\mathbf{h}$ and $\mathbf{h}'$ is defined as

$$s(h, h') = \sum_{k=1}^{m} I(\mathbf{h}[k], \mathbf{h}'[k]), \tag{2}$$

where $I(a, b) = 0$ if $a = b$ and $I(a, b) = 1$ otherwise. As the hamming distance is length-dependent, we define the following $A(\mathbf{h}, \mathbf{h}')$ as a length-independent distance between haplotype-alleles $\mathbf{h}$ and $\mathbf{h}'$:

$$A(\mathbf{h}, \mathbf{h}') = \frac{s(\mathbf{h}, \mathbf{h}')}{m}. \tag{3}$$

2.2. *HIT algorithm*

The Haplotype Inference Technique (HIT) algorithm (Rastas et al., 2005) is an HMM-based algorithm for phasing unphased-diplotypes. The algorithm utilizes the HMM (Rabiner and Juang, 1986). The HMM of the HIT is designed to simulate multiple set of ancestors (i.e., founders).[3] The HMM is trained from a set

[3] According to Rastas et al. (2005), the optimal number of ancestors is around 7 for most cases. Thus, we also use the HMM model with 7 ancestors in the experiments in Section 4.

of unphased-diplotypes in an unsupervised way with the EM algorithm (Durbin et al., 1998). Figure 1 shows the HMM model used in the HIT. The HIT algorithm phases an unphased haplotype-diplotype by heuristically finding the haplotype-diplotype with the highest emission probability from the HMM.

2.3. Clustering methods

In this section, we describe the clustering method and the method for evaluating the results, which we will use in Section 4.

2.3.1. Ward's method. We use Ward's minimum variance algorithm (Team RDC, 2007; Ward, 1963; Ward and Hook, 1963), which is a widely used hierarchical clustering method, to infer clusters based on the ASD or the HHD in Section 4.[4] Given n items $I_1, I_2, \cdots, I_n$, a distance matrix $\{w_{ij}\}$ where w_{ij} denotes the distance between I_i and I_j, and some fixed positive integer k ($k < n$), the Ward's method clusters the n items into k clusters by the following $n - k - 1$ steps.[5] At first the algorithm considers n clusters each of which contains only 1 item, i.e., $\mathcal{C}_1 = \{\{I_1\}, \{I_2\}, \cdots, \{I_n\}\}$. Then the algorithm reduces the number of clusters one by one in each step as follows. In the m-th step of the algorithm, two clusters are merged into a cluster to minimize $\sum_{C \in \mathcal{C}_{m+1}} \sum_{I_i, I_j \in C} w_{ij}^2 / |C|$, where $\mathcal{C}_i$ denotes the set of clusters before the i-th step of the algorithm. This bottom-up approach is repeated until $|\mathcal{C}_m| = k$.

2.3.2. How to evaluate the clustering results. To evaluate the clustering results, we use the classification error rate (CER) (Gao and Starmer, 2007). The CER is the rate of elements that are assigned to incorrect clusters in clustering results. To know the assignment is correct or not, we need to know the labels of each cluster, but Ward's algorithm does not assign any labels onto the output clusters. In the experiment, we use the minimum CER among all the possible assignments of the population labels, to evaluate the clustering results.

3. NEW UNPHASED-DIPLOTYPE DISTANCE MEASURES

In this section, we first propose in Section 3.1 a new measure for the distance between two unphased-diplotypes, the PMD. The PMD is a general concept of distance measures, and we will give an example of the PMD which we call the HHD in Section 3.2. In Section 3.3, we discuss the properties of the proposed measures.

3.1. Population model–based distance

Before defining our new measure called the PMD, we first extend the haplotype similarity measure described in Section 2.1.2 so that we can deal with the distances between two haplotype-diplotypes instead of haplotype-alleles, as follows. Let $a = \{\mathbf{h}_1, \mathbf{h}_2\}$ and $a' = \{\mathbf{h}'_1, \mathbf{h}'_2\}$ be haplotype-diplotypes to be compared, where $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}'_1, \mathbf{h}'_2 \in \mathcal{S}^m$. We define the distance between haplotype-diplotypes a and a' as

$$H(a, a') = \min\left\{\frac{A(\mathbf{h}_1, \mathbf{h}'_1) + A(\mathbf{h}_2, \mathbf{h}'_2)}{2}, \frac{A(\mathbf{h}_1, \mathbf{h}'_2) + A(\mathbf{h}_2, \mathbf{h}'_1)}{2}\right\}, \tag{4}$$

where A is the haplotype similarity measure defined in Section 2.1.2. But we cannot compute this value for unphased-diplotypes, as we cannot know the actual haplotype-diplotypes. To enable it, we extend the above haplotype-diplotype distance H for unphased-diplotypes by utilizing some given population model $\mathcal{M}$ as follows.

For any unphased-diplotype, we can enumerate corresponding haplotype-diplotype candidates.[6] For example, there are four haplotype-diplotype candidates for unphased-diplotype $\{1, 0\} - \{1, 0\} - \{1, 0\}$, i.e., $\{111, 000\}$, $\{110, 001\}$, $\{101, 010\}$, and $\{011, 011\}$. For unphased-diplotypes $\mathbf{g}, \mathbf{g}' \in \mathcal{D}^m$, let $c_i = \{\mathbf{h}_{i1}, \mathbf{h}_{i2}\}$ ($1 \le i \le M$) and $c'_j = \{\mathbf{h}'_{j1}, \mathbf{h}'_{j2}\}$ ($1 \le j \le M'$) be the i-th and the j-th candidate haplotype-diplotypes for

[4] We used the statistical software, R, to implement this algorithm.

[5] The ASD or the HHD values will be used as w_{ij} in Section 4.

[6] *Phasing* is the process of finding the most probable haplotype-diplotype, utilizing some population information.

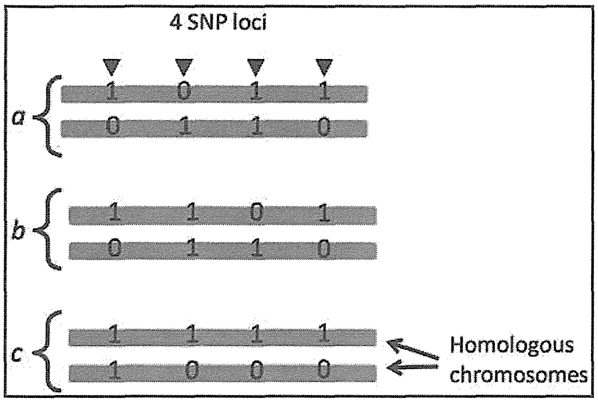

FIG. 2. Haplotype-diplotype examples on which we can observe difference between the ASD and the PMD.

g and **g**$'$, respectively. M and M' are the numbers of haplotype-diplotype candidates for **g** and **g**$'$, respectively.

If we were given a population model $\mathcal{M}$, we can compute the probability $Prob(c|\mathbf{g}, \mathcal{M})$ that a haplotype-diplotype candidate c is correct for the unphased-diplotype data **g**. Let $p_i = Prob(c_i|\mathbf{g}, \mathcal{M})$ and $p'_j = Prob(c'_j|\mathbf{g}', \mathcal{M})$ be the conditional probabilities of the candidate haplotype-diplotypes c_i and c'_j under the model $\mathcal{M}$. Then the $PMD_{\mathcal{M}}$ between two haplotype-diplotypes **g** and **g**$'$ is defined as follows:

$$PMD_{\mathcal{M}}(\mathbf{g}, \mathbf{g}') = \sum_{i=1}^{M}\sum_{j=1}^{M'} H(c_i, c'_j) \cdot q_i \cdot q'_j, \tag{5}$$

where $q_i = p_i / (\sum_{k=1}^{M} p_k)$ and $q'_j = p'_j / (\sum_{k=1}^{M'} p'_k)$. q_i and q'_j are the normalized predicted conditional probabilities of the candidate haplotype-diplotypes c_i and c'_j, respectively.[7] Note that the PMD is the expected value of the distance between candidate haplotype-diplotypes, $H(c_i, c'_j)$, under the population model$\mathcal{M}$.

3.2. *HIT HMM-based Distance*

To compute the PMD in Section 3.1, we need an appropriate model for the population. In the following, we propose an example of the PMD that we call the HHD.[8] To define the HHD, we propose to use the HMM model used in the HIT algorithm (Rastas et al., 2005) (described in Section 2.2) as the population model for the PMD as follows.

The HMM defined in the HIT algorithm can be considered as a predicted population model. Thus, we first train the HMM from all the unphased-diplotype data that are in our hand, and then we define the HHD as follows. Let $\mathcal{M}^*$ denote the HMM model obtained with the HIT. Then we define the HHD as

$$HHD(\mathbf{g}, \mathbf{g}') = PMD_{\mathcal{M}^*}(\mathbf{g}, \mathbf{g}'). \tag{6}$$

Note that the probability of each haplotype-diplotype candidate is computed as the conditional emission probability of the candidate from the HMM, which can be computed by the forward algorithm (Durbin et al., 1998) for the HMM.

3.3. *Discussions on the PMD*

3.3.1. The PMD and the multiple founder hypothesis. In many regions (especially in important regions) of the human genome, the haplotype-alleles of the majority in populations can be categorized into a small number of types (Bhatia et al., 2010; Cirulli and Goldstein, 2010), which suggest that only a small number of founder (or ancestral) haplotype-alleles spread over the population on those regions. This

[7] Note that $\sum_{k=1}^{M} p_k = \sum_{k=1}^{M} p'_k = 1$ and there is no need to normalize the probabilities if we enumerate all the candidates. But we need to normalize them in case we ignore the candidates with very small probabilities. When we compute the HHD (which will be introduced in Section 3.2), we ignore candidates with very small probabilities.

[8] We also introduce other simpler examples of the PMD in Section 3.3.1.

TABLE 1. DISTANCES BETWEEN THE INDIVIDUALS IN FIGURE 2

	(1) ASD				(2) $H = PMD_{\mathcal{M}_1}$				(3) $PMD_{\mathcal{M}_2}$		
	a	*b*	*c*		*a*	*b*	*c*		*a*	*b*	*c*
a	0	0.25	0.25	*a*	0	0.25	0.5	*a*	0	0.301	0.450
b	—	0	0.25	*b*	—	0	0.5	*b*	—	0	0.500
c	—	—	0	*c*	—	—	0	*c*	—	—	0

hypothesis of the existence of (a few but) multiple founder haplotype-alleles is very important and effective for various kinds of research, for example, the design of the experiments of linkage disequilibrium mapping (Chung et al., 2008; Gonzalez et al., 1999; Haiman et al., 2003) and the evolutionary history analysis of populations (Ahmad et al., 2002; Gaudieri et al., 1997).

The PMD well reflects the existence of the founder haplotype-alleles. In the example given in Figure 2, there are three individuals with haplotype-diplotypes $a = \{1011, 0110\}$, $b = \{1101, 0110\}$, and $c = \{1111, 1000\}$, but we assume that we know only the unphased-haplotypes, i.e., $\{1, 0\} - \{1, 0\} - \{1, 1\} - \{1, 0\}$, $\{1, 0\} - \{1, 1\}$-$\{1, 0\} - \{1, 0\}$ and $\{1, 1\} - \{1, 0\} - \{1, 0\} - \{1, 0\}$, respectively. We can easily see that the ASD between any two of these three individuals is 0.25 (Table 1*(1)*), and therefore we cannot cluster these three individuals based on the ASD.

The distance between two sequences are often measured by the number of point mutations between them (i.e., we consider two sequences to be very distant to each other if there are many mutations between them). We can define the number of mutations under the assumption of existence of multiple founder haplotype-alleles (for details, see the Appendix). Table 2 shows the number under the assumption that there are two founder haplotype-alleles. According to the table, the clustering result of the three individuals should be the one in Figure 3, which cannot be obtained with the ASD. Note that the clustered individuals *a* and *b* share the same haplotype-allele, i.e., 0110, which also supports the validity of the clustering result.

Unlike the ASD, the haplotype-diplotype distance *H* reflects the numbers in Table 2 very well. The *H* value between individuals *a* and *b* is 0.25, which is the same value as the ASD, but *H* between *a* and *c* and *H* between *b* and *c* are 0.5 (Table 1*(2)*), which enable us to cluster the individuals as in Figure 3. It means the *H* values are more appropriate than the ASD values under the existence of the founder haplotype-alleles, at least in this case.

But we cannot compute the real *H* values unless we know the real haplotype-diplotypes. Instead, we can estimate them by computing the PMD if we are given some population model. Consider the two population models given in Table 3, where haplotype frequencies in the population are given.[9] Under the model $\mathcal{M}_1$, we can phase any of the three individuals' unphased-haplotypes correctly with 100% confidence, and the resulting $PMD_{\mathcal{M}_1}$ values are the same as the *H* values (Table 1*(2)*). But we cannot predict unphased-haplotypes with such high confidence in many cases, as in the case of the population model $\mathcal{M}_2$ where we have multiple haplotype-diplotype candidates for each unphased diplotype (see Table 4 and Table 1*(3)*).

If we cluster the three individuals based on the $H = PMD_{\mathcal{M}_1}$ values, we can obtain the same clusters as in Figure 3. Furthermore, we can still get the same clusters even if we use the $PMD_{\mathcal{M}_2}$ values instead. Thus, we assume that the PMD is more suitable than the ASD under the multiple founder hypothesis, if we are given an appropriate population model.

3.3.2. Influences of the linkage equilibrium. It is easy to imagine that the linkage equilibrium (LE) and the linkage disequilibrium (LD) should affect the similarity measures. In fact, the variance of the distribution of the ASD values among the individuals should converges to some value in $\Theta(1/m)$ where *m* is the number of the SNP loci in the region according to the central limit theorem, if the loci are independent to each other. It means that the variance of the ASD values should be smaller on the regions of LE. The PMD and its example HHD should also be influenced by the LE/LD. We compared the influences of the LE/LD to the ASD and the HHD by checking distances on the LE/LD regions obtained from the HapMap database (release 24) (International HapMap Consortium, 2005) as follows.

[9] The population models could be represented by many other methods. For example, we consider HMM-based models in Section 3.2.

TABLE 2. NUMBER OF MUATIONS BETWEEN EACH INDIVIDUAL UNDER THE ASSUMPTION THAT THERE ARE TWO FOUNDERS

	a	*b*	*c*
a	0	2	4
b	—	0	4
c	—	—	0

See Appendix how we obtain the number of mutaions for each pair of individuals.

We can determine whether a region is near to LE or to LD by counting the number of haplotype tagging SNPs (htSNPs) (Carlson et al., 2004; Johnson et al., 2001; Ke and Cardon, 2003; Meng et al., 2003; Rinaldo et al., 2005). The htSNPs are selected so that each SNP in the given region has a correlation larger than a threshold with at least one of the htSNPs. Thus, the regions with many htSNPs can be considered to be near the LE, and regions with few htSNPs can be considered to be near the LD.

We divided the set of SNPs in chromosome 1 into 658 blocks, each of which consists of 100 consecutive SNPs. For each block B, we counted the number h_B of htSNPs obtained by the software Tagger (de Bakker et al., 2005) with the default settings. We selected 100 blocks with the 100 smallest h_B values as the LD regions and also selected 100 blocks with the 100 largest h_B values as the LE regions.

For each of all these regions, we computed the ASD and the HHD measures among the 270 individuals in HapMap (which are the same as the 270 individuals used in Section 4), and computed the variances among the obtained $270 \times 269/2 = 36315$ distances of the ASD and of the HHD. Table 5 shows the difference between the variances of the ASD and the HHD measures. According to the P-values in the table, the HHD reflects the LD/LE effects more than the ASD.

4. APPLICATION TO HAPMAP DATA SETS

4.1. Data sets

In the experiments in Section 4.2, we will use the unphased-diplotype data sets of 22 autosomal chromosomes and X chromosome derived from HapMap release 24 (International HapMap Consortium, 2005). The data sets consist of unphased-diplotypes of 270 individuals: 90 Yoruba in Ibadan, Nigeria (YRI); 90 Utah residents with ancestry from northern and western Europe (CEU, from the CEPH diversity panel); and 90 Japanese in Tokyo, Japan, and Han Chinese in Beijing, China (CHB + JPT). There are 894,398 SNPs that are genotyped for all the above 270 individuals, which we used for our experiments. We divided the SNP set into 8,930 blocks, each of which consists of consecutive 100 SNPs, and we will perform comprehensive experiments against each of these blocks in Section 4.2.

4.2. Experimental results

In this section, we demonstrate the impact of incorporating the population information, by comparing the clustering accuracies by the ASD and that by the HHD on the HapMap data described in Section 4.1.

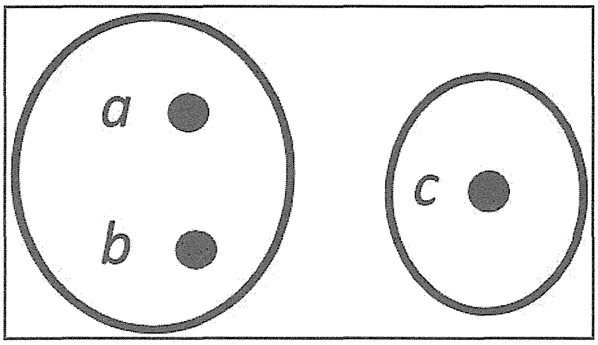

FIG. 3. Clustering results for individuals in Figure 2 based on the numbers of mutations (Table 2), $H = PMD_{\mathcal{M}_1}$ distances (Table 1*(2)*), or $PMD_{\mathcal{M}_2}$ distances (Table 1*(3)*). On the other hand, the ASD distances (Table 1*(1)*) cannot deduce this result.

Table 3. Population Model Examples Given as Haplotype-Allele Frequencies

	Frequency in population	
Haplotype-allele	*(i)* $\mathcal{M}_1$	*(ii)* $\mathcal{M}_2$
1111	0.40	0.20
1110	0.00	0.07
1101	0.20	0.08
1011	0.25	0.10
0011	0.00	0.05
0110	0.10	0.30
0101	0.00	0.05
1100	0.00	0.05
1000	0.05	0.10
Others	0.00	0.00

Against each of the 8,930 blocks, we performed Ward's clustering algorithm (see Section 2.3.1) based on the ASD and also did the same based on the HHD, and compared the CERs (see Section 2.3.2) of their results (Table 6). The difference of the results in relation to the number of htSNPs, i.e., h_B (see Section 3.3.2), is also shown.

The mean of CERs based on the HHD (i.e., 0.3557) is better than that for the ASD (i.e., 0.3611). The P-value of the t-test to see the difference between them is 0.004177, which means the CERs of the HHD is significantly better than that of the ASD. The number of data sets where the HHD (or the ASD) shows better performance than the ASD (or the HHD) are checked with the sign test. Among all the data sets, the HHD is superior to the ASD on 4366 data sets and inferior to the ASD in 3696 data sets. The results of two measures were the same in the other 868 data sets. The P-value of the sign test of all of these results is $8.98 \cdot 10^{-14}$, which means that the HHD is significantly superior to the ASD.

The CERs decrease with increasing h_B for both the ASD and the HHD, but the differences of CERs between the ASD and the HHD also increases as h_B increase (Fig. 4). We call the result HDD's success if the HHD's CER is lower than that of the ASD, and vice versa. The ratio of the HHD's success increases with increasing h_B. The ratio of ASD's success also increases with increasing h_B. The difference of ratios of success between the ASD and the HHD is getting larger as h_B increases. The ratio of the case when the ASD and the HHD have the same results are getting lower as h_B increases (Fig. 5).

The HHD is superior to ASD especially when $80 \leq h_B < 90$. It is a reasonable result as we should be able to better cluster individuals if we have more information (i.e., LE). The difference of ratios of success

Table 4. Conditional Probabilities of Candidate Haplotype-Diplotypes for Individuals in Figure 2 Based on the Population Models in Table 3

			Conditional probability	
Individual	*Unphased-diplotype*	*Candidate haplotype-diplotype*	*(i)* $\mathcal{M}_1$	*(ii)* $\mathcal{M}_2$
	{1,0}-{1,0}-{1,1}-{1,0}	{1011, 0110}	1.0000	0.8955
a		{1110, 0011}	0.0000	0.1045
		Others	0.0000	0.0000
	{1,0}-{1,1}-{1,0}-{1,0}	{1101, 0110}	1.0000	0.8727
b		{1110, 0101}	0.0000	0.1273
		Others	0.0000	0.0000
	{1,1}-{1,0}-{1,0}-{1,0}	{1111, 1000}	1.0000	0.8000
c		{1011, 1100}	0.0000	0.2000
		Others	0.0000	0.0000

TABLE 5. MEANS OF VARIANCES OF ASD/HHD MEASURES ON THE REGIONS WHERE THE SNPs ARE WEAKLY CORRELATED AND HIGHLY CORRELATED IN CHROMOSOME 1

	Mean of variances		
	LE	*LD*	P-*value*
ASD	0.00267	0.00546	$2.066 \cdot 10^{-16}$
HHD	0.00248	0.00539	$1.637 \cdot 10^{-17}$

The LE and LD columns show the means of variances on the LE regions (i.e., regions with many htSNPs) and those on the LD regions (i.e., regions with a few htSNPs), respectively. The difference of the variances between weakly and highly correlated regions are tested by t-test for each of the measures. The P-value column shows the P-value of the t-test.

between the ASD and the HHD also becomes largest when $80 < h_B < 90$. In this case, the HHD is superior on 13 data sets, while the ASD is superior only on six data sets among the remaining 18 data sets.

5. CONCLUSION

We proposed a new inter-diplotype similarity measure that we call the PMD. The PMD improves the previous ASD measure by utilizing a population model. As one of such population models, we propose to use the HMM population model used in the phasing algorithm HIT. We call the PMD based on the HIT's HMM the HHD. The HHD utilizes the predicted conditional probabilities of haplotype-diplotypes of unphased-diplotype emitted from the HIT's HMM. Based on comprehensive experiments over 8930 genome-wide data sets of HapMap, we showed that the HHD significantly outperforms the ASD. We also discussed the relationships between the clustering accuracies and the LD.

There are many future tasks to do related to this work. The HHD requires much larger computation time than the ASD, and one future task should be to improve the computation speed of the HHD. There are still data sets for which the HHD is not superior to the ASD. It would be very interesting if we can predict the regions where the HHD is inferior to the ASD, before computing these measures. Another future task is to improve the population model, as it should directly improve the performance of the PMD. From the biological viewpoint, it would also be very interesting if we can utilize our clustering algorithms to identify

TABLE 6. THE EXPERIMENTAL RESULTS AND THEIR RELATIONSHIPS TO THE H_B VALUES

		Mean of CERs		*Comparison of CERs*			
h_B	*♯blocks*	*ASD*	*HHD*	$CER_{ASD} < CER_{HHD}$	$CER_{HHD} < CER_{ASD}$	$CER_{ASD} = CER_{HHD}$	P-*value of sign test*
0 ∼ 10	1	0.5630	0.5630	0 (0.0)	0 (0.0)	1 (1.0)	
10 ∼ 20	44	0.4733	0.4678	9 (0.2045)	13 (0.2955)	22 (0.5)	0.5235
20 ∼ 30	223	0.4363	0.4305	62 (0.2780)	82 (0.3677)	79 (0.3543)	0.1130
30 ∼ 40	993	0.4240	0.4207	380 (0.3827)	418 (0.4209)	195 (0.1964)	0.1902
40 ∼ 50	2364	0.3929	0.3877	975 (0.4124)	1131 (0.4784)	258 (0.1091)	$7.276 \cdot 10^{-4}$*
50 ∼ 60	3063	0.3567	0.3514	1327 (0.4332)	1528 (0.4989)	208 (0.06793)	$1.808 \cdot 10^{-4}$*
60 ∼ 70	1822	0.3052	0.2997	772 (0.4237)	970 (0.5324)	80 (0.04391)	$2.303 \cdot 10^{-6}$*
70 ∼ 80	399	0.2584	0.2465	165 (0.4135)	211 (0.5288)	23 (0.05764)	0.02018*
80 ∼ 90	21	0.2178	0.1944	6 (0.2857)	13 (0.6190)	2 (0.09524)	0.1671
90 ∼ 100	0	—	—	—	—	—	—
Total	8930	0.3611	0.3557	3696 (0.4139)	4366 (0.4889)	868 (0.09720)	$8.98 \cdot 10^{-14}$*

The ♯blocks column shows the numbers of blocks with the specified h_B values. In the Comparison of CERs columns, the $CER_{ASD} < CER_{HHD}/CER_{ASD} > CER_{HHD}/CER_{ASD} = CER_{HHD}$ columns show the numbers (and the ratios) of data (with the specified h_B values) where the ASD performed better/the HHD performed better/the performance of the two measures are exactly the same, respectively. $x \sim y$ indicates that $x \leq h_B < y$, and * means the result of the sign test is significant (i.e., ≤ 0.05).

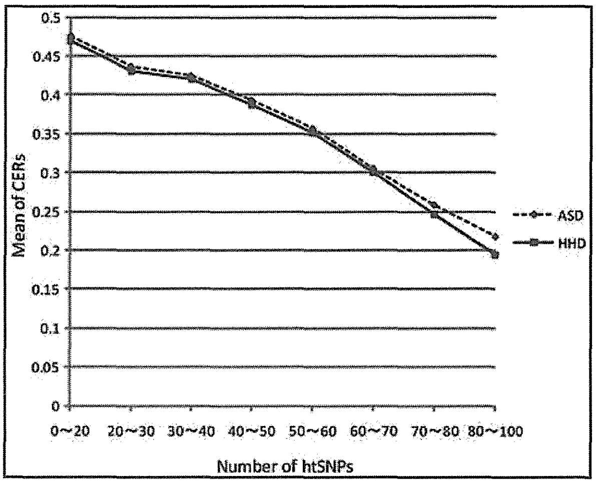

FIG. 4. The plot of h_B values and the means of CERs for both the ASD and the HHD. $x \sim y$ indicates that $x \leq h_B < y$. The HHD is superior to the ASD in all the cases.

gene functions of the target genome regions, especially the regions that affect the disease prevalence and drug responses (Bamshad et al., 2004; Wiencke, 2004; Wilson et al., 2001).

6. APPENDIX

Counting number of mutations under founder hypothesis

Suppose that founder haplotype-alleles $\mathbf{f}_1, \ldots, \mathbf{f}_m$ has been evolved into the present-day haplotpye-alleles of individuals p and q, without any recombinations. Let p_1 and p_2 be the haplotype-alleles of p and $\mathbf{q}_1$ and $\mathbf{q}_2$ be the haplotype-alleles of q. We can consider that the number of mutations between p and q under the assumption of founders $\mathbf{f}_1, \ldots, \mathbf{f}_m$ as

$$S_{\mathbf{f}_1,\ldots,\mathbf{f}_m}(p,q) = \min\left\{ \sum_{i=1}^{2} \min_{j=1}^{m}\{dist(\mathbf{p}_i,\mathbf{f}_j) + dist(\mathbf{q}_i,\mathbf{f}_j)\},\right.$$
$$\left. \sum_{i=1}^{2} \min_{j=1}^{m}\{dist(\mathbf{p}_i,\mathbf{f}_j) + dist(\mathbf{q}_{2-i},\mathbf{f}_j)\} \right\}, \qquad (7)$$

where *dist*() denotes the ordinary number of mutations between the two sequences.

But we cannot know the appropriate set of founder haplotype-alleles. Instead, we can define the number of mutations between two individuals under the assumption that there are m founders as

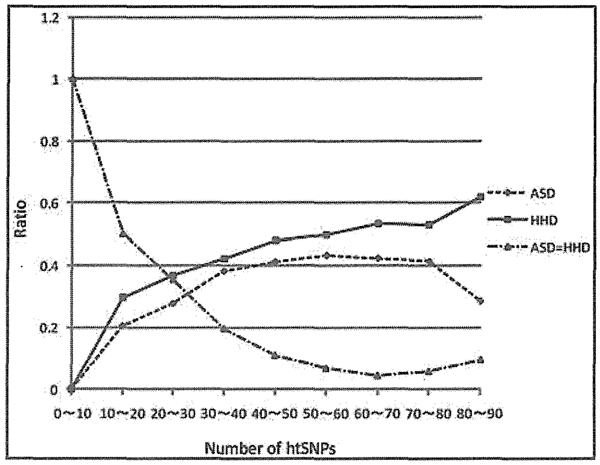

FIG. 5. The plot of h_B values and the ratios of success for both the ASD and the HHD. The line ASD = HHD indicates the results in which the performance of the two measures are the exactly the same. $x \sim y$ indicates that $x \leq h_B < y$. The HHD is superior to the ASD in all the cases.

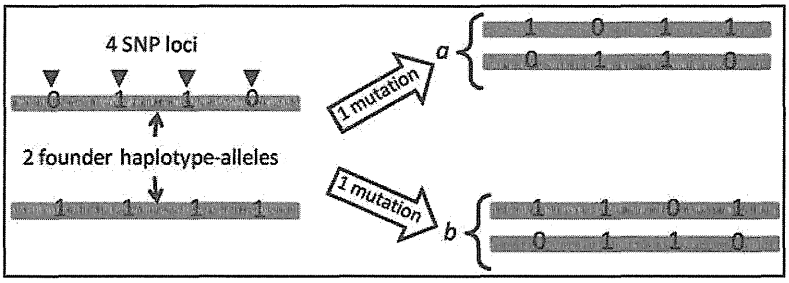

FIG. 6. The optimal founder haplotype-allele pair (when $m = 2$) for the individuals a and b in Figure 2.

$$S^*_m(p, q) = \min_{\mathbf{f}_1, \ldots, \mathbf{f}_m} S_{\mathbf{f}_1, \ldots, \mathbf{f}_m}(p, q). \tag{8}$$

Table 2 shows all the $S^*_2()$ values for all the pairs among individuals a, b, and c in Figure 2. Figure 6 shows the founder pair $\mathbf{f_1}$, $\mathbf{f_2}$ that minimizes the $S_{\mathbf{f}_1, \mathbf{f}_2}(a, b)$ value.

ACKNOWLEDGMENTS

The experiments in this work were done on the Super Computer System of the Human Genome Center, the Institute of Medical Science, the University of Tokyo.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Ahmad, T., Neville, M., Marshall, S.E., et al. 2002. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Hum. Mol. Genet. 12, 647–656.

Bamshad, M., Wooding, S., Salisbury, B.A., et al. 2004. Deconstructing the relationship between genetics and race. Nat. Rev. Genet. 5, 598–609.

Beaty, T.H., Fallin, M.D., Hetmanski, J.B., et al. 2005. Haplotype diversity in 11 candidate genes across four populations. Genetics 171, 259–267.

Bhatia, G., Bansal, V., Harismendy, O., et al. 2010. A covering method for detecting genetic associations between rare variants and common phenotypes. Plos Comput. Biol. 6, 1–12.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., et al. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457.

Carlson, C.S., Eberle, M.A., Rieder, M.J., et al. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74, 106–120.

Chung, P.Y.J., Beyens, G., Guanabens, N., et al. 2008. Founder effect in different European countries for the recurrent P392L SQSTM1 mutation in Paget's disease of bone. Calcif. Tissue. Int. 83, 34–42.

Cirulli, E.T., and Goldstein, D.B. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11, 415–425.

Conrad, D.F., Jakobsson, M., Coop, G., et al. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 38, 1251–1260.

Cornuet, J.M., Sylvain, P., Luikart, G., et al. 1999. New methods employing multilocus genotypes to select or exlcude populations as origins of individuals. Genetics 153, 1989–2000.

Cover, T.M., and Thomas, J.A. 1991. *Elements of Information Theory*, John Wiley & Sons, New York.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., et al. 2005. Efficiency and power in genetic association studies. Nat. Genet. 37, 1217–1223.

Durbin, R., Eddy, S., Krogh, A., et al. 1998. *Biological Sequence Analysis*. Cambridge Press, New York.

Ester, M., Kriegel, H.P., Sander, J., et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining* 226–231.

Fallin, D., Cohen, A., Essioux, L., et al. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res. 11, 143–151.

Falush, D., Stephens, M., and Pritchard, J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587.

Gao, X., and Martin, E.R. 2009. Using allele sharing distance for detecting human population stratification. Hum. Hered. 68, 3.

Gao, X., and Starmer, J. 2007. Human population structure detection via multilocus genotype clustering. BMC Genet. 8, 34.

Gaudieri, S., Leelayuwat, C., Tay, G.K., et al. 1997. The Major Histocompatability Complex (MHC) contains conserved polymorphic genomic sequences that are shuffled by recombination to form ethnic-specific haplotypes. J. Mol. Evol. 45, 17–23.

Gonzalez, E., Bamshad, M., Sato, N., et al. 1999. Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. Proc. Natl. Acad. Sci. USA 96, 12004–12009.

Haiman, C.A., Stram, D.O., Pike, M.C., et al. 2003. A comprehensive haplotype analysis of CYP19 and breast cancer risk: The Multiethnic Cohort. Hum. Mol. Genet. 12, 2679–2692.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437, 1299–1320. Available at www.hapmap.org. Accessed November 1, 2011.

Isaev, A. 2004. *Introduction to mathematical methods to bioinformatics*. Springer, New York.

Jakobsson, M., Scolz, S.W., Scheet, P., et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451, 998–1003.

Jin, L., Zhu, W., and Guo, J. 2010. Genome-wide association studies using haplotype clustering with a new haplotype similarity. Genet. Epidemiol. 34, 633–641.

Johnson, G.C.L., Esposito, L., Barratt, B.J., et al. 2001. Haplotype tagging for the identification of common disease genes. Nat. Genet. 29, 233–237.

Kaufman, L, and Rousseeuw, P. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

Ke, X., and Cardon, L.R. 2003. Efficient selective screening of haplotype tag SNPs. Bioinformatics 19, 287–288.

Kim, S., and Misra, A. 2007. SNP genotyping: technologies and biomedical applications. Annu. Rev. Biomed. Eng. 9, 289–320.

Lesk, A.M. 2005. Introduction to Bioinformatics, 2nd ed. *Oxford*, New York.

Li, J., and Jiang, T. 2005. Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics 21, 4384–4393.

Li, J., Zhou, Y., and Elston, R.C. 2006. Haplotype-based quantitative trait mapping using a clustering algorithm. BMC Bioinform. 7, 258.

Mao, X., Bigham, A.W., Mei, R., et al. 2007. A genomewide admixture mapping panel for Hispanic/Latino populations. Am. J. Hum. Genet. 80, 1171–1178.

Meng, Z., Zaykin, D.V., Xu, C., et al. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am. J. Hum. Genet. 73, 115–130.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rabiner, L.R., and Juang, B.H. 1986.An introduction to hidden Markov models. *IEEE ASSP Mag*. 3, 4–16.

Rastas, P., Koivisto, P.M., Mannila, H., et al. 2005. A hidden Markov technique for haplotype reconstruction. *Lect. Notes Bioinform.* 3692, 140–151.

Rinaldo, A., Bacanu, S., Devlin, B., et al. 2005. Characterization of multilocus linkage disequilibrium. Genet. Epidemiol. 28, 193–206.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Small, K.M., Mialet-Perez, J., Seman, C.A., et al. 2004. Polymorphisms of cardiac presynaptic α_{2C} adrenergic receptors: diverse intragenic variability with haplotype-specific functional effects. Proc. Natl. Acad. Sci. USA 101, 13020–13025.

Team RDC. 2007. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*.

Tzeng, J.Y., Devlin, B., Wasserman, L., et al. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am. J. Hum. Genet. 72, 891–902.

Ward, J.H. 1963. Hierarchical grouping procedure to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Ward, J.H., and Hook, M.E. 1963. Application of an hierarchical grouping procedure to a problem of grouping profiles. Educ. Psychol. Measure. 23, 69–81.

Wiencke, J.K. 2004. Impact of race/ethnicity on molecular pathways in human cancer. Nat. Rev. Cancer 4, 79–84.

Wilson, F.W., Weale, M.E., Smith, A.C., et al. 2001. Population genetic structure of variable drug response. Nat. Genet. 29, 265–269.

Witherspoon, D.J., Wooding, S., Rogers, A.R., et al. 2007. Genetic similarities within and between human populations. Genetics 176, 351–359.

Yang, Y., and Tabus, I. 2007. Haplotype block partitioning using a normalized maximum likelihood model. *Proc. IEEE Genomic Signal Process. Stat.* 1–4.

Address correspondence to:
Ms. Ritsuko Onuki
Bioinformatics Center
Institute for Chemical Research
Kyoto University
Gokasho, Uji
Kyoto 611-0011, Japan

E-mail: onuki@hgc.jp

Clinical Infectious Diseases Advance Access published September 24, 2012

MAJOR ARTICLE HIV/AIDS

Single Nucleotide Polymorphisms in *ABCC2* Associate With Tenofovir-Induced Kidney Tubular Dysfunction in Japanese Patients With HIV-1 Infection: A Pharmacogenetic Study

Takeshi Nishijima,[1,2] Hirokazu Komatsu,[3] Koichiro Higasa,[4] Misao Takano,[1] Kiyoto Tsuchiya,[1] Tsunefusa Hayashida,[1] Shinichi Oka,[1,2] and Hiroyuki Gatanaga[1,2]

[1]AIDS Clinical Center, National Center for Global Health and Medicine, Tokyo; [2]Center for AIDS Research, Kumamoto University; [3]Department of Community Care, Saku Central Hospital, Nagano; and [4]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Japan

Background. Tenofovir is a widely used antiretroviral drug although it can cause kidney tubular dysfunction (KTD). The aim of this study was to determine the association between polymorphisms in genes encoding drug transporters and KTD in Japanese patients treated with tenofovir.

Methods. The association between tenofovir-induced KTD and 14 single nucleotide polymorphisms (SNPs) in the *ABCC2*, *ABCC4*, *ABCC10*, *SCL22A6*, and *ABCB1* genes was investigated in 190 Japanese patients. KTD was diagnosed by the presence of at least 3 abnormalities in the following parameters: fractional tubular resorption of phosphate, fractional excretion of uric acid, urinary β2-microglobulin, urinary α1-microglobulin, and urinary N-acetyl-β-D-glucosaminidase. Genotyping was performed by allelic discrimination using TaqMan 5′-nuclease assays with standard protocols. Associations between genotypes and KTD were tested by univariate and multivariate logistic regression analyses.

Results. KTD was diagnosed in 19 of the 190 (10%) patients. Univariate and multivariate analyses showed a significant association between KTD and genotype CC at position −24 CC (adjusted odds ratio [OR], 20.08; 95% confidence interval [CI], 1.711–235.7; $P = .017$) and genotype AA at position 1249 (adjusted OR, 16.21; 95% CI, 1.630–161.1; $P = .017$) of *ABCC2*. Multivariate analysis showed higher adjusted OR for patients with both homozygotes (adjusted OR, 38.44; 95% CI, 2.051–720.4; $P = .015$). *ABCC2* haplotype −24T and 1249G was a protective haplotype for KTD (OR, 0.098; 95% CI, .002–.603; $P = .003$

Conclusions. This is the first study of our knowledge to identify the association between SNPs in *ABCC2* and tenofovir-induced KTD in an Asian population. Close monitoring of renal function is warranted in tenofovir-treated patients with these SNPs.

Tenofovir disoproxil fumarate (TDF), a prodrug of tenofovir, is a nucleotide reverse transcriptase inhibitor widely used for the treatment of human immunodeficiency virus type 1 (HIV-1) infection and hepatitis B infection [1–4]. Tenofovir is excreted by a combination of glomerular filtration and active tubular secretion. Although the nephrotoxicity of tenofovir is regarded mild and tolerable [5–7], several cases of tenofovir-induced nephrogenic diabetes insipidus, Fanconi syndrome, and acute renal failure have been reported, and prognosis of renal function with long-term tenofovir use remains unknown [8–10].

The mechanism of tenofovir-induced kidney damage is not fully understood. However, mitochondrial damage in the proximal renal tubular cells was observed in patients with prominent tenofovir-induced kidney tubular dysfunction (KTD) [11, 12].

Received 20 May 2012; accepted 24 August 2012.

Correspondence: Hiroyuki Gatanaga, MD, PhD, AIDS Clinical Center, National Center for Global Health and Medicine, 1-21-1, Toyama, Shinjuku, Tokyo 162-0052, Japan (higatana@acc.ncgm.go.jp).

Clinical Infectious Diseases

DOI: 10.1093/cid/cis772

Because the characteristics and severity of tenofovir-induced KTD vary widely among individuals, the role of host genetics has drawn a particular attention. Single nucleotide polymorphisms (SNPs) in transporter proteins of renal tubular cells have been investigated to elucidate their roles in tenofovir-induced KTD [13–15].

Tenofovir enters kidney tubular cells through the basolateral membrane and is transported mainly by organic anion transporter (OAT) 1 and, to a lesser extent, OAT 3, encoded by genes *SLC22A6* and *SLC22A8*, respectively [16]. Tenofovir is excreted into the urine at the apical membrane by 2 transporters on the luminal membrane; multidrug resistance protein (MRP) 4 and MRP 2, encoded by the adenosine triphosphate–binding cassette (ABC) genes *ABCC4* and *ABCC2*, respectively [17, 18]. Although the role of MRP4 in transporting tenofovir has been well established, that of MRP 2 remains controversial [19, 20]. Recently, MRP 7, encoded by *ABCC10* gene, was also reported to take part in the excretion of tenofovir [21]. P-glycoprotein is a membrane protein expressed on the cells of renal proximal tubule, intestine, and hepatocytes. Encoded by *ABCB1* gene, P-glycoprotein transports TDF, the prodrug of tenofovir. SNPs on *ABCB1* might alter the expression of P-glycoprotein and thus affect exposure of tenofovir [22–24].

Previous studies reported inconsistent findings on the association of the SNPs of the transporter protein on tenofovir-induced KTD [13–15]. Several pathological processes could induce KTD, such as active infection, inflammation, diabetic nephropathy, concurrent use of nephrotoxic drugs, and preexisting renal impairment, and thus it is difficult to evaluate KTD induced exclusively by tenofovir [25]. Moreover, drug interaction with other antiretrovirals, especially ritonavir-boosted protease inhibitors, modifies tenofovir clearance and thus the severity of tenofovir-induced KTD [26, 27]. Previous studies examined patients treated with various antiretroviral combinations, which might also contribute to the inconsistent findings. Thus, the effect of SNPs on tenofovir-induced KTD remains to be clarified and isolated from other abovementioned conventional risk factors for KTD [15, 28]. Of note, the population investigated in previous studies on the role of SNPs in tenofovir-induced KTD was mostly whites, and patients of other genetic background have hardly been examined.

Based on the above background, the present study was designed to elucidate the association between polymorphisms in genes encoding drug transporters in renal tubular cells and tenofovir-induced KTD, in a setting designed to exclude other predisposing or intervening factors: the inclusion of Japanese patients with HIV infection on the same antiretroviral combination with suppressed HIV-1 viral load, and free of preexisting renal impairment, major comorbidities, and active infections.

METHODS

Ethics Statement

This study was approved by the Human Genetics Research Ethics Committee of the National Center for Global Health and Medicine, Tokyo, Japan. Each patient included in this study provided a written informed consent for genetic testing and publication of clinical data for research purposes. The study was conducted according to the principles expressed in the Declaration of Helsinki.

Study Design

We performed a single-center cohort study to cross-sectionally elucidate the association between SNPs in genes encoding renal tubular transporters in Japanese patients with HIV infection and tenofovir-induced KTD.

Study Subjects

The study included consecutive Japanese patients with HIV infection, aged >17 years, with HIV-1 viral load <200 copies/mL, and on at least 4-week treatment with once-daily ritonavir (100 mg)–boosted darunavir (800 mg) plus fixed dose tenofovir (300 mg)/emtricitabine (200 mg), seen at our clinic between 1 October 2011 and 31 March 2012. The exclusion criteria were (1) active infection, (2) malignancy, (3) diabetes mellitus, defined by the use of anti-diabetic agents or fasting plasma glucose >126 mg/dL or plasma glucose >200 mg/dL on two different days, (4) alanine aminotransferase 2.5 times more than the upper limit of normal, (5) estimated glomerular filtration rate (eGFR) calculated by Cockcroft-Gault equation of <50 mL/minutes [creatinine clearance = [(140 − age) × weight (kg)]/(serum creatinine × 72)(×0.85 for females)] [29], and (6) patients without consent to the study.

Measurements

Blood and spot urine samples were collected either on the day of enrollment or on the next visit, together with body weight measurement. The blood samples were used to measure serum creatinine, serum uric acid, serum phosphate, CD4 count, and C-reactive protein, whereas urine samples were used to measure phosphate, uric acid, creatinine, β2-microglobulin (β2M), α1-microglobulin (α1M), and N-acetyl-β-D-glucosaminidase (NAG). The values of β2M, α1M, and NAG measured in the urine samples were expressed relative to urinary creatinine of 1 g/L (/g Cr).

Urinary concentrations of β2M and α1M were measured with latex aggregation assay kits (β2M: BMG-Latex X1 "Seiken"; Denka Seiken Co, Niigata, Japan; α1M: Eiken α1M-III; Eiken Chemical Co, Tokyo, Japan), and those of NAG by colorimetric assay of enzyme activity with 6-methyl-2-pyridyl-N-acetyl-1-thio-β-D-glucosaminide as substrate (Nittobo Medical Co, Tokyo).

Definition of Renal Proximal Tubular Dysfunction

KTD was defined as the presence of at least 3 abnormalities in the following 5 parameters: fractional tubular resorption of phosphate {1 − [(urine phosphate × serum creatinine)/(urine creatinine × serum phosphate)]} × 100 of <82%, fractional excretion of uric acid {[(urine uric acid × serum creatinine)/(urine creatinine × serum uric acid)] × 100)} of >15%, β2-microglobulinuria (β2M > 1000 μg/g Cr), α1-microglobulinuria (α1M > 16.6 mg/g Cr), and high-NAG level in urine (NAG > 5.93 U/g Cr). The above cutoff levels were selected on the basis of data reported previously by various investigators [15, 30, 31].

The potential risk factors for KTD were determined according to previous studies and collected together with the basic demographics from the medical records [6, 27, 32, 33]. They included age, sex, body weight, and presence or absence of other medical conditions (concurrent use of nephrotoxic drugs such as ganciclovir, sulfamethoxazole/trimethoprim, and nonsteroidal antiinflammatory agents, coinfection with hepatitis B, defined by positive hepatitis B surface antigen, coinfection with hepatitis C, defined by positive HCV viral load, hypertension, defined by current treatment with antihypertensive agents or 2 successive measurements of systolic blood pressure >140 mmHg or diastolic blood pressure >90 mmHg at the clinic, dyslipidemia, defined by current treatment with lipid-lowering agents or 2 successive measurements of either low-density lipoprotein cholesterol >140 mg/dL, high-density lipoprotein cholesterol <40 mg/dL, total cholesterol >240 mg/dL, triglyceride >500 mg/dL. At our clinic, blood pressure and body weight are measured every visit. We used the data on or closest to and preceding the day of blood/urine sample collection by no more than 180 days.

Genetic Polymorphisms

SNPs in genes encoding tubular transporters were selected on the basis of their functional significance, findings of previously published reports, and/or reported minor-allele frequencies >5% in the Japanese [13–15, 21, 28]. The allele frequency data for the Japanese were obtained from the Japanese Single Nucleotide Polymorphisms (JSNP) database [34]. The 14 SNPs selected were (1) *ABCC2* (encodes MRP2) −24C → T (in the promoter; rs717620); 1249G → A (Val417Ile; rs2273697); 2366C → T (Ser789Phe; rs56220353); 2934G → A (Ser978Ser; rs3740070), (2) *ABCC4* (encodes MRP4) 559G → T (Gly187Trp; rs11568658); 912G → T (Lys304Asn; rs2274407); 2269G → A (Glu757Lys; rs3765534); 3348A → G (Lys1116Lys; rs1751034); 4135T → G [in the 3′ untranslated region (UTR); (rs3742106)]; 4976T → C (3′ UTR; rs1059751), (3) *ABCC10* (encodes MRP10) 526G → A (intron; rs9349256); 2759T → C (Ile920Thr; rs2125739), (4) *SLC22A6* (encodes OAT1) 180C → T (Asn60Asn; rs11568630), and (5) *ABCB1* (encodes P-glycoprotein) 2677T → A/G (A:Ser893Thr, G:Ser893Ala; rs2032582).

Pharmacogenetic Analyses

Genomic DNA was extracted from peripheral-blood leukocytes using the protocol described in the sheet enclosed with the QIAamp DNA MiniKit (Qiagen, Valencia, California). All genotyping was performed by allelic discrimination using TaqMan 5′-nuclease assays with standard protocols (TaqMan SNP Genotyping Assays; Applied Biosystems, Foster City, California). The primer and probe sequences are available on request.

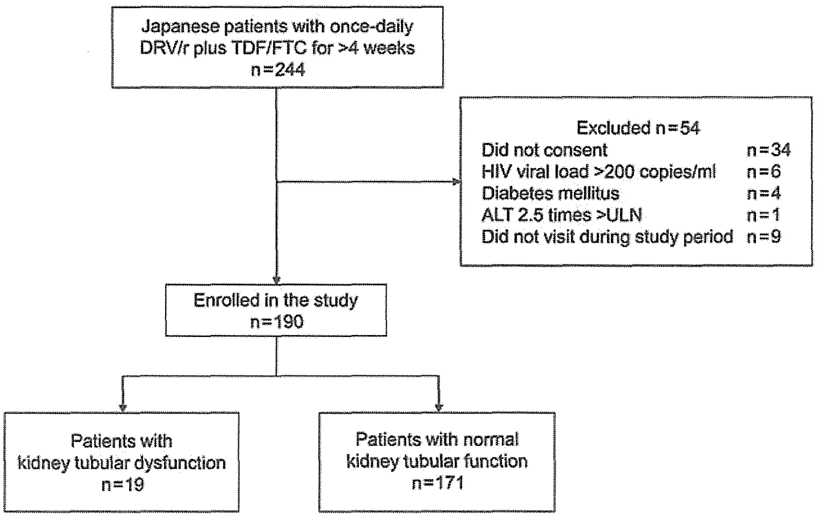
Japanese patients with once-daily DRV/r plus TDF/FTC for >4 weeks n=244
Excluded n=54
Did not consent n=34
HIV viral load >200 copies/ml n=6
Diabetes mellitus n=4
ALT 2.5 times >ULN n=1
Did not visit during study period n=9
Enrolled in the study n=190
Patients with kidney tubular dysfunction n=19
Patients with normal kidney tubular function n=171

Figure 1. Patient enrollment. Abbreviations: ALT, alanine transaminase; DRV/r, ritonavir-boosted darunavir; HIV, human immunodeficiency virus; TDF/FTC, tenofovir/emtricitabine; ULN, upper limit of normal.

Statistical Analysis

Baseline characteristics were compared between patients with KTD and without tubular dysfunction by the Student *t* test for continuous variables and by either the χ^2 test or Fisher exact test for categorical variables. Statistical comparisons for genotype frequencies between 2 groups were made by use of 2 × 3 table Fisher exact test (2 × 6 table for rs2032582). Associations between genotypes and KTD were tested by univariate and multivariate logistic regression analyses. The impact of other variables was estimated with univariate analysis, and those with $P < .20$ were incorporated into multivariate analysis, in addition to the basic demographics such as age and sex. Statistical significance was defined at 2-sided *P* value < .05. We used odds ratios (ORs) and 95% confidence intervals (95% CIs) to estimate the impact of each variable on KTD. The Haploview software was used to test Hardy-Weinberg equilibrium and *ABCC2* and *ABCC4* haplotype analysis. All other statistical analyses were performed with the Statistical Package for Social Sciences ver. 17.0 (SPSS, Chicago, Illinois).

RESULTS

A total of 190 patients who provided blood and urine samples and satisfied the inclusion and exclusion criteria were enrolled in the study (Figure 1). KTD was diagnosed in 19 of the 190 patients (10%). The baseline characteristics and laboratory data for patients with and without KTD are listed in Table 1. Patients with KTD were older ($P < .001$), had smaller body weight ($P = .006$) and lower eGFR ($P = .003$), and were more likely to be hypertensive than patients with normal tubular function ($P = .088$). The median duration of tenofovir therapy was 71.5 weeks (interquartile range [IQR]: 36.8–109.2 weeks) for the entire study population, which was not different between the 2 groups ($P = .888$).

Table 1. Characteristics of Patients With and Without Kidney Tubular Dysfunction

	Patients With KTD (n = 19)	Patients With Normal Tubular Function (n = 171)	*P* Value
Variables for kidney tubular markers			
Urinary β2M (μg/g Cr)[a]	3066 (2247–10068)	209.2 (114.2–536.2)	<.001
Urinary α1M (mg/g Cr)[a]	26.5 (19.8–37.4)	7.95 (5.02–11.9)	<.001
Urinary NAG (U/g Cr)[a]	9 (6.2–14.3)	3.74 (2.84–4.95)	<.001
Fractional tubular resorption of phosphate[a]	83.9 (81.7–92)	91.9 (88.8–94.4)	<.001
Fractional excretion of uric acid[a]	9.7 (8.1–12.4)	6.4 (5.0–9.0)	<.001
Contribution of each parameter to KTD			
Urinary β2M > 1000 μg/g Cr, No. (%)	19 (100)	21 (12.3)	<.001
Urinary α1M > 16.6 mg/g Cr, No. (%)	18 (94.7)	17 (9.9)	<.001
Urinary NAG >5.93 U/g Cr, No. (%)	17 (89.5)	23 (13.5)	<.001
Fractional tubular resorption of phosphate <82%, No. (%)	5 (26.3)	2 (1.2)	<.001
Fractional excretion of uric acid >15%, No. (%)	2 (10.5)	4 (2.3)	.112
Characteristics			
Sex (male), No. (%)	18 (94.7)	166 (97.1)	.473
Age[a]	60 (41–62)	38 (32–42)	<.001
Route of transmission (homosexual contact), No. (%)	16 (84.2)	153 (89.5)	.528
Weight (kg)[a]	56 (53.5–66.5)	67.2 (58.1–75)	.006
Estimated glomerular filtration rate (mL/minutes/1.73 m^2)[a]	75.5 (62.8–93.5)	87.7 (77.5–98)	.003
Serum creatinine (mg/dL)[a]	0.85 (0.68–0.96)	0.80 (0.73–0.88)	.168
CD4 cell count (μL)[a]	380 (194–501)	379 (275–533)	.261
Serum phosphate (mg/dL)[a]	3.4 (2.7–3.7)	3.2 (2.9–3.6)	.815
Serum uric acid (mg/dL)[a]	4.7 (4.2–5.7)	5.6 (4.8–6.4)	.080
Nephrotoxic drug, No. (%)	2 (10.5)	12 (7.0)	.420
Hepatitis C, No. (%)	0 (0)	3 (1.8)	.728
Hepatitis B, No. (%)	2 (10.5)	24 (14)	.501
Dyslipidemia, No. (%)	4 (21.1)	54 (31.6)	.253
Hypertension, No. (%)	8 (42.1)	42 (24.6)	.088
C-reactive protein (mg/dL)[a]	0.07 (0.03–0.28)	0.07 (0.03–0.16)	.277
Duration of treatment with TDF (weeks)[a]	60.3 (17.7–115.4)	73.3 (37.7–109.1)	.888

Abbreviations: KTD, kidney tubular dysfunction; NAG, N-acetyl-β-D-glucosaminidase; TDF, tenofovir disoproxil fumarate.

[a] Median (interquartile range).

Table 2. Genotype Frequencies at *ABCC2*, *ABCC4*, *ABCC10*, *SLC22A6*, and *ABCB1* in Patients With and Without Kidney Tubular Dysfunction

Genotype	Amino Acid	Patients With KTD (n = 19)	Patients With Normal Tubular Function (n = 171)	*P* Value[a]
ABCC2 (MRP2)				
−24 C → T, rs717620				
C/C		18 (94.7)	108 (63.2)	
C/T		1 (5.3)	52 (30.4)	.018
T/T		0 (0)	11 (6.4)	
1249 G → A, rs2273697	Val417Ile			
G/G		11 (57.9)	133 (77.8)	
A/G		5 (26.3)	34 (19.9)	.017
A/A		3 (15.8)	4 (2.3)	
2366 C → T, rs56220353	Ser789Phe			
C/C		19 (100)	167 (97.7)	
C/T		0 (0)	3 (1.8)	1.000
T/T		0 (0)	1 (0.6)	
2934 G → A, rs3740070	Ser978Ser			
G/G		18 (94.7)	159 (93.0)	
G/A		1 (5.3)	11 (6.4)	1.000
A/A		0 (0)	1 (0.6)	
ABCC4 (MRP4)				
559 G → T, rs11568658	Gly187Trp			
G/G		13 (68.4)	133 (77.8)	
G/T		4 (21.1)	34 (19.9)	.126
T/T		2 (10.5)	4 (2.3)	
912G → T, rs2274407				
G/G		13 (68.4)	102 (59.6)	
T/G		6 (31.6)	52 (30.4)	.461
T/T		0 (0)	17 (9.9)	
2269 G → A, rs3765534	Glu757Lys			
G/G		15 (78.9)	129 (75.4)	
G/A		2 (10.5)	35 (20.5)	.241
A/A		2 (10.5)	7 (4.1)	
3348 A → G, rs1751034	Lys1116Lys			
A/A		13 (68.4)	98 (57.3)	
A/G		3 (15.8)	58 (33.9)	.185
G/G		3 (15.8)	15 (8.8)	
4135 T → G, rs3742106				
T/T		6 (31.6)	46 (26.9)	
T/G		7 (36.8)	79 (46.2)	.707
G/G		6 (31.6)	46 (26.9)	
4976T → C, rs1059751				
T/T		6 (31.6)	46 (26.9)	
T/C		5 (26.3)	86 (50.3)	.090
C/C		8 (42.1)	39 (22.8)	
ABCC10 (MRP7)				
526G → A, rs9349256				
G/G		4 (21.1)	32 (18.7)	
A/G		9 (47.4)	65 (38)	.569
A/A		6 (31.6)	74 (43.3)	

Table 2 continued.

Genotype	Amino Acid	Patients With KTD (n = 19)	Patients With Normal Tubular Function (n = 171)	*P* Value[a]
2759T → C, rs2125739	Ile920Thr			
T/T		15 (71.4)	131 (77.5)	
T/C		6 (28.6)	31 (18.3)	.488
C/C		0 (0)	7 (4.1)	
SLC22A6 (OAT1)				
180C → T, rs11568630				
C/C		18 (94.7)	164 (95.9)	
C/T		1 (5.3)	7 (4.1)	.577
T/T		0 (0)	0 (0)	
ABCB1 (P-glycoprotein)				
2677T → A/G, rs2032582	A:Ser893Thr G:Ser893Ala			
T/T		0 (0)	47 (27.5)	
T/A		3 (15.8)	14 (8.2)	
G/G		4 (21.1)	36 (21.1)	.002
G/T		8 (42.1)	46 (26.9)	
G/A		1 (5.3)	24 (14)	
A/A		3 (15.8)	4 (2.3)	

Abbreviation: KTD, kidney tubular dysfunction.

[a] By Fisher exact test.

Table 2 summarizes the distribution of genotypes at the *ABCC2*, *ABCC4*, *ABCC10*, *SLC22A11*, and *ABCB1* genes in the 2 groups. All polymorphisms were in Hardy-Weinberg equilibrium with a cutoff *P* value of .001. In single SNP analysis, a higher percentage of patients with KTD were found among genotype CC at position −24 and genotype AA at position 1249 of *ABCC2*, compared to patients with other genotypes (−24 CC; 14.3% [in 18 of 126 patients] vs 1.6% [in 1 of 64 patients]; *P* = .004) (1249 AA; 42.9% [in 3 of 7 patients] vs 8.7% [in 16 of 183 patients]; *P* = .023), respectively. The percentage of patients with KTD was also higher among genotype AA at position 2677 of *ABCB1*, compared to patients with other genotypes (2677 AA; 42.9% [in 3 of 7 patients] vs. 8.7% [in 16 of 183 patients]; *P* = .023). KTD was marginally associated with genotype AA at position 559 and genotype GG at position 4976 of *ABCC4* (*P* = .112, and .090, respectively).

Association of Genotypes with KTD

Univariate analysis showed a significant association between KTD and patients with genotype CC at position −24 (OR, = 10.50; 95% CI, 1.369–80.55; *P* = .024) and patients with genotype AA at position 1249 (OR, 7.828; 95% CI, 1.609–38.10; *P* = .011) of *ABCC2* (Table 3). The risk for KTD was higher in patients with both genotype CC at position −24 and genotype AA at position 1249 (OR, 31.88; 95% CI, 3.131–324.5; *P* = .003). Genotype AA at position 2677 of *ABCB1* was also significantly associated with KTD (OR, 7.828; 95% CI, 1.609–38.10; *P* = .011). Furthermore, old age (per 1 year, OR, 1.165; 95% CI, 1.100–1.233; *P* < .001), low body weight (per 1 kg decrement, OR, 1.076; 95% CI, 1.021–1.135; *P* = .007), and low eGFR (per 1 mL/minutes/1.73 m^2 decrement, OR, 1.052; 95% CI, 1.016–1.090; *P* = .004) were also associated with KTD.

Multivariate analysis identified genotype CC at position −24 and genotype AA at position 1249 of *ABCC2* as independent risks for KTD after adjustment for sex, age, weight, eGFR, and hypertension (adjusted OR, = 20.08; 95% CI, 1.711–235.7; *P* = .017) (adjusted OR, 16.21; 95% CI, 1.630–161.1; *P* = .017), respectively (Table 4). Patients with both of the abovementioned two homozygotes showed higher adjusted OR in multivariate analysis (adjusted OR, 38.44; 95% CI, 2.051–720.4; *P* = .015) (Table 4). On the other hand, genotype AA at position 2677 of *ABCB1* was not significantly associated with KTD in multivariate analysis adjusted for the abovementioned variables (adjusted OR, 1.686; 95%CI, .163–17.43; *P* = .661).

Association of Haplotypes at ABCC2 and ABCC4 with KTD

Haplotype construction was performed with the 4 identified SNPs with *P* < .10 in univariate analysis: *ABCC2*, −24 C → T, 1249 G → A; *ABCC4*, 559 G → T, 4976T → C (Table 4). Haplotypes with frequency of >1% were analyzed. *ABCC2* haplotype CA was significantly associated with TDF-induced KTD (OR, 2.910; 95% CI, 1.295–6.221; *P* = .011), whereas *ABCC2* haplotype TG was a protective haplotype (OR, 0.098; 95% CI, .002–.603; *P* = .003). *ABCC4* haplotype TT was marginally