Journal of Human Genetics (2010) 55, 500–506

www.nature.com/jhg

ORIGINAL ARTICLE

Making a haplotype catalog with estimated frequencies based on SNP homozygotes

Yumi Yamaguchi-Kabata[1], Tatsuhiko Tsunoda[2], Atsushi Takahashi[1], Naoya Hosono[3], Michiaki Kubo[3], Yusuke Nakamura[4,5] and Naoyuki Kamatani[1]

Understanding the structure and frequencies of haplotypes is important for associating genetic polymorphisms with a given trait and for inferring the genetic genealogy of alleles in a population. Single nucleotide polymorphism (SNP) haplotypes can be determined without ambiguity when an individual does not have more than one heterozygous site in a given genomic region. Using genome-wide SNP genotypes for 3397 individuals from the Japanese population, we detected SNP homozygotes in the genomic regions of 1955 genes, determined haplotypes, and examined the efficiency of haplotype frequency estimation based on the proportion of SNP homozygotes in the sample. The estimated haplotype frequencies were very similar to the frequencies obtained by two statistical methods, PHASE and SNPHAP. We applied this approach to the genomic regions of 11 351 genes, and the results suggested that the sum of the frequencies of unobserved haplotypes is negligible for an analysis of a 100 kb genomic region with ~20 SNPs. Determination of haplotypes from homozygotes using genotype data from thousands of individuals, without a long computation time, appears to be useful for detecting real haplotypes including some low-frequency haplotypes. In addition, the unambiguously determined haplotypes with their estimated frequencies can be used as a catalog of haplotypes for the population, which is useful for the design of genome-wide association studies.
Journal of Human Genetics (2010) **55,** 500–506; doi:10.1038/jhg.2010.56; published online 20 May 2010

Keywords: haplotype; haplotype frequency; homozygotes; linkage disequilibrium; single-nucleotide polymorphisms

INTRODUCTION

In a genome-wide association study (GWAS), one tries to find the genomic loci responsible for a given trait. This is performed by finding polymorphic markers that are associated with the trait. Recently, single-nucleotide polymorphism (SNP) markers (landmark SNPs) are extensively used for this purpose. In such studies, much effort has been directed at understanding haplotype structure and haplotype frequency.[1–4]

Haploid human genomes differ by one SNP in every 1200–1700 bp.[5,6] Even though significant progress has been made in sequencing technologies, determination of haplotypes by sequencing is expensive and time consuming. Another approach is to infer haplotypes by statistical methods from genotype data. A pioneering algorithm to infer haplotypes using genotype data by Clark,[7] which starts with detection of homozygous individuals, used a parsimonious algorithm and heuristic search. Many statistical methods for inferring haplotypes or estimating haplotype frequency were subsequently developed,[8–16] and some of them are widely used to infer haplotypes from SNP genotype data.

After a dense SNP map on the human genome became available,[5,17–19] several studies examined the extent of linkage disequilibrium.[1,2,4] The SNP discovery project, one of the Japan Millennium Genome Projects, identified over 170 000 SNPs by sequencing gene regions in the human genome of DNA samples from 24 individuals.[19,20] Subsequently, genotype data were obtained for a larger number of individuals from the Japanese population, and allele frequencies for common SNPs were estimated (http://snp.ims.u-tokyo.ac.jp/). We built a genome-wide map of linkage disequilibrium for the gene-based SNPs,[2] and this led to GWASs as early as 2002–2003.[21,22] The International HapMap project[23,24] provided genome-wide SNP genotype data for several selected populations. In that study, haplotype inference was conducted with the genotype data by statistical methods.[23] The HapMap project also provided guidelines for selecting tag SNPs for association studies and has led to an increase in the number of GWASs.[25] Statistical approaches for inferring haplotypes are quite accurate for haplotypes that exist at high frequencies. However, such methods usually predict too many haplotypes in which frequencies for some haplotypes are very low, and it is

[1]Laboratory for Statistical Analysis, Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), Minato-ku, Tokyo, Japan; [2]Laboratory for Medical Informatics, Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), Tsurumi-ku, Yokohama, Japan; [3]Laboratory for Genotyping Development, Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), Tsurumi-ku, Yokohama, Japan; [4]Laboratory for Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo, Japan and [5]Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), Tsurumi-ku, Yokohama, Japan
Correspondence: Dr Y Yamaguchi-Kabata, Laboratory for Statistical Analysis, Research Group for Medical Informatics, Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan.
E-mail: yyamaguc@src.riken.jp
Received 11 January 2010; revised 12 April 2010; accepted 27 April 2010; published online 20 May 2010

usually difficult to tell which of these haplotypes really exist. Furthermore, the accuracy of haplotype inference from data of unrelated individuals was found to be lower than the haplotype inference from family data.[26] Another approach to determining haplotypes in a population is to use hemizygous tissues in which haplotypes are unambiguously determined. For example, complete hydatidiform moles (CHMs), which have chromosomes from a single sperm, have been used to determine the haplotypes of the Japanese population.[27]

Clark[7] showed that homozygous segments without multiple heterozygous sites in an individual can be used to define haplotypes. Searching runs of SNP homozygosity for autozygous genomic regions are useful for mapping recessive diseases.[28–32] Using genotype data from thousands of individuals in the BioBank Japan Project,[33] we attempted to (1) determine haplotypes by detecting SNP homozygotes for genomic regions of interest and (2) estimate haplotype frequencies based on the proportion of homozygous individuals in the sample by assuming that genotype frequencies are in Hardy–Weinberg equilibrium. To determine the accuracy and efficiency of estimating haplotype frequencies from homozygous individuals, we analyzed SNP genotype data for 3397 individuals from the Japanese population.

First, we conducted a test analysis with 'definitive haplotypes' from 74 complete hydatidiform mole samples in the Kyushu University Definitive Haplotype Database (D-haploDB),[27,34] and compared the estimated haplotype frequencies with those obtained by two statistical methods, PHASE and SNPHAP, to examine the reliability of haplotype determination and haplotype frequency determination. Second, we applied this approach to the genomic regions of all human genes to evaluate the efficiency of the method in various conditions.

MATERIALS AND METHODS

Determination of haplotypes and estimation of haplotype frequencies

Using SNP genotype data from a sufficient number of individuals, we examined the efficiency of haplotype determination from homozygous individuals for a given range of genomic region and estimation of haplotype frequencies from the proportion of the homozygous individuals in the sample. First, we select homozygous individuals whose SNP genotypes were homozygous for all the SNPs in a given genomic region. Then, we determine haplotypes assuming that two copies of the same haplotype are present for each homozygous individual. We estimate frequency of each haplotype from the proportion of homozygous individuals for the haplotype in the sample. Suppose we have a sample of N diploid individuals from a population. Let f $(=f_1, \ldots, f_M)$ denotes the set of haplotype frequencies for the observed haplotypes (arbitrarily labeled 1, …, M). Assuming the Hardy–Weinberg equilibrium, the expected number of homozygotes for a particular haplotype i, n_i, is Nf_i^2. The frequency of haplotype i can be estimated as follows,

$$\hat{f}_i = \sqrt{p_i}$$

where p_i $(=n_i/N)$ denotes the proportion of the homozygotes of haplotype i in the sample.

Subjects and genotype data

Genotype data were obtained from 3397 self-identified Japanese individuals. These individuals consists of healthy controls from Midosuji Rotary Club and case individuals for 13 of the 47 diseases that are studied in the BioBank Japan Project.[33] All the patients provided written informed consent to participate in the BioBank Japan Project. The BioBank Japan Project was approved by the ethical committees at The Institute of Medical Science, The University of Tokyo, and the Center for Genomic Medicine (formerly, SNP Research Center), Institutes of Physical and Chemical Research (RIKEN).

All the Japanese DNA samples were genotyped for 568 666 SNPs by using Illumina 550K arrays (Illumina, San Diego, CA, USA). Genotyped SNPs in autosomes (chromosomes 1–22) were selected for further analyses if they satisfy both of the following two criteria: (1) call rates were high enough ($\geqslant$99%), (2) no abnormality was detected by visual inspection of raw data of genotyping when there was a departure from Hardy–Weinberg's equilibrium of genotype frequencies ($P<10^{-6}$; $10^{-6}\leqslant P<10^{-3}$ and call rate <0.9998). After the selection of SNPs, the genotype data for 547 458 SNPs were used in further analyses.

Test analysis with the definitive haplotypes

To compare the determined haplotypes and their frequencies by different approaches, we selected 79 007 SNPs that were genotyped in both of the 547 458 SNPs (genotyped by the Illumina 550K arrays) and 81 250 SNPs (genotyped by the Perlegen platform) in the definitive haplotypes from 74 CHMs in D-haploDB (http://orca.gen.kyushu-u.ac.jp/)[27,34] (Supplementary Figure 1). After discarding monomorphic SNPs in the 3397 individuals, 79 005 SNPs were used for the following analyses.

Data of genomic locations for the RefSeq genes (transcripts) were retrieved from the Entrez Gene website (http://www.ncbi.nlm.nih.gov/gene) in the NCBI bioinformatics resources. Genomic regions between the start and end positions for each transcript were selected for analysis. We selected genomic regions having at least three SNPs for the haplotype analysis.

Comparison of haplotype frequencies with other approaches

Haplotypes for the analyzed regions were inferred and their frequencies were estimated by using computer programs, PHASE (http://stephenslab.uchicago.edu/software.html)[9] and SNPHAP (http://www-gene.cimr.cam.ac.uk/clayton/software/). All of the 3397 individuals were included in the analysis. The haplotype frequencies estimated from homozygotes were compared with the haplotype frequencies estimated by the two statistical methods, PHASE and SNPHAP, and the correlation coefficient of haplotype frequencies by two approaches was calculated. We also normalized the haplotype frequencies from the homozygotes so that the sum of frequencies of the observed haplotypes equaled 1.0, and the haplotype frequencies were compared with those obtained by the two statistical methods in the same way.

Similarly, the haplotype frequencies were also compared with those of the 'definitive haplotypes' in 74 CHMs in D-haploDB.[27,34] Genotype data for the definitive haplotypes from the 74 CHMs were downloaded from D-haploDB,[27,34] and the list of haplotypes for the 1955 genomic regions were generated from the genotype data.

Application to genomic regions of all human genes

We used the genotype data for genome-wide 546 457 SNPs that satisfy quality control filters (see above), and selected 404 758 SNPs by discarding SNPs whose minor allele frequencies were less than 0.05 (Supplementary Figure 3). The genomic region for each transcript included an additional 3000 bp region for both 5′ and 3′ side of the transcript. We selected 11 351 genomic regions which have at least three SNPs for the haplotype analysis.

RESULTS

Test analysis with statistical approaches and the definitive haplotypes

To examine the efficiency of the haplotype analysis from SNP homozygotes by comparing with different approaches, we selected 79 005 SNPs, which were genotyped in both the 3397 Japanese individuals in the BioBank Japan Project[33] and the 74 CHM samples.[27] Using genomic locations for human transcripts, we selected the genomic regions for genes that had at least three of these SNPs. For the three or more SNPs in each genomic region, the average distance between SNPs was 20 215 bp. The minor allele frequencies ranged from 0.02 to 0.50, and the average was 0.271.

We selected 1955 genomic regions with 3–10 analyzed SNPs for further analyses. In these genomic regions, we detected homozygous individuals in which genotypes were homozygous for all the analyzed SNPs. The proportion of homozygotes in the sample was 0.37 on average for the 741 regions having three SNPs, and decreased with increasing number of SNPs (Table 1). In total, we detected

Table 1 Detection of homozygous individuals in the 3397 individuals for the 1955 genomic regions

No. of SNPs	*No. of regions*	*No. of homozygous individuals*[a]
3	741	1256.3 (0.3698)
4	412	1003.5 (0.2954)
5	269	739.4 (0.2177)
6	191	637.6 (0.1877)
7	105	441.2 (0.1299)
8	92	444.0 (0.1307)
9	85	301.4 (0.0887)
10	60	273.9 (0.0806)

Abbreviation: SNP, single-nucleotide polymorphism.
[a]The analyzed genomic regions with 3–10 SNPs were divided according to the number of SNPs, and the average numbers and proportions (in parentheses) of homozygous individuals between the regions with the same number of SNPs were calculated.

Table 2 Number of haplotypes detected in homozygous individuals and predicted by statistical approaches

			Average number of haplotypes		
			Deterministic approach	*Statistical approach*	
No. of SNPs (L)	2^L[a]	*No. of regions*	*Homozygotes*	*PHASE*[b]	*SNPHAP*[b]
3	8	741	4.86	4.93 (1.75)	4.94 (1.62)
4	16	412	6.91	6.99 (4.30)	7.06 (4.17)
5	32	269	9.81	10.12 (9.09)	10.19 (8.93)
6	64	191	12.12	12.17 (17.21)	12.30 (17.52)
7	128	105	15.82	16.16 (32.87)	16.23 (33.21)
8	256	92	18.03	—	—
9	512	85	20.28	—	—
10	1024	60	21.60	—	—

Abbreviation: SNP, single-nucleotide polymorphism.
[a]The number of possible haplotypes with each number of SNPs.
[b]Haplotypes whose frequencies were less than 0.01 were counted separately, and the average numbers of those haplotypes ($0.0001 <$ frequency < 0.01) are shown in parentheses.

17 739 haplotypes for the 1955 genomic regions by detecting individuals in which genotypes were homozygous for all the analyzed SNPs. The numbers of haplotypes whose frequencies are higher than 0.01 were very similar to those obtained by the two statistical methods (PHASE and SNPHAP) (Table 2). However, the statistical methods predicted a large number of haplotypes whose frequencies are less than 0.01. Similarly, the numbers of haplotypes in the 74 CHMs in D-haploDB were very similar to those detected in homozygotes for genomic regions having 3–5 SNPs (Supplementary Table 1). However, as the number of the SNPs in the regions increased, the number of haplotypes in the 74 CHMs became larger than the number of haplotypes detected in homozygotes in the 3397 individuals (Supplementary Table 1).

Frequencies of the detected haplotypes were estimated from the proportion of homozygotes in the sample. The lowest haplotype frequency was 0.0172 (the square root of 1/3397), a frequency that was observed for 3633 haplotypes in 1118 genes. The highest haplotype frequency (0.946) was observed in the *INVS* gene for haplotype GGC (rs2787366, rs1999877, rs2787390), for which 3041 of 3397 individuals were homozygous.

The haplotype frequencies estimated from the proportions of homozygotes were highly correlated with the results obtained by the two statistical methods: the correlation coefficients were 0.9986 ($P < 2.2\times10^{-16}$) for PHASE (Figure 1a) and 0.9985 ($P < 2.2\times10^{-16}$) for SNPHAP (Figure 1c). The correlation coefficient of the haplotype frequencies with the definitive haplotypes from 74 CHMs was 0.9691 ($P < 2.2\times10^{-16}$) (Figure 1e). Although this is a comparison of haplotype frequencies between the SNP homozygotes in the 3397 individuals and the 74 CHMs from the Japanese population, the haplotype frequencies in the two samples were very similar. When the haplotype frequencies were normalized so that their sum equaled 1.0, they were still highly correlated (r=0.9983, $P < 2.2\times10^{-16}$) with the frequencies obtained by the statistical methods (Figure 1b and d) and with those in D-haploDB (r=0.9688, $P < 2.2\times10^{-16}$, Figure 1f), although the correlations were slightly weaker than those obtained without the normalization.

For the genomic regions having three polymorphic SNPs, the average number of haplotypes identified by our method that were also among the definitive haplotypes from the 74 CHMs in D-haploDB was 4.7 (Supplementary Figure 2). This suggests that this study independently identified most of the definitive haplotypes in the 74 CHMs. On the other hand, 280 haplotypes in 741 genomic regions were present only in the 74 CHMs in D-haploDB, and 142 haplotypes were present only in our result (Supplementary Figure 2).

Application to genomic regions for all human genes

The efficiency of detection of haplotypes in a region from SNP homozygotes depends on several factors such as the number of SNPs in the region, the length of the region, the level of linkage disequilibrium and the selection of SNPs. In particular, as the number of SNPs increases, both the number of possible haplotypes and the number of actual haplotypes would increase, and the actual haplotypes may contain haplotypes that exist at very low frequencies. On the other hand, the proportion of SNP homozygotes would decrease as the number of SNPs increases. Generating a list of haplotypes from SNP homozygotes with a larger number of SNPs may not include the haplotypes that exist at very low frequencies. Therefore, we applied our approach to a larger number of genomic regions and evaluated the results according to the number of SNPs and the level of linkage disequilibrium.

We focused on the genomic regions for all the human transcripts (see Materials and methods and Supplementary Figure 3) and used genotype data for 404 758 SNPs after discarding SNPs, whose minor allele frequencies were less than 0.05 (see Materials and methods). Then, we selected 11 351 genomic regions which have at least three SNPs. Although the number of analyzed SNPs on those regions ranged from 3 to 706, the number of the analyzed SNPs were less than 20 for a majority of the regions analyzed (9335/11 351, data not shown). For the three or more SNPs in each genomic region, the average distance between SNPs was 5290 bp, and the average minor allele frequency was 0.27. The estimated frequencies of haplotypes were compared with those based on SNPHAP (Supplementary Figure 4). The haplotype frequencies estimated from the proportions of homozygotes were highly correlated (r=0.9991, $P < 2.2\times10^{-16}$) with the results obtained by the SNPHAP program.

The proportions of SNP homozygotes (+ marks in Figure 2) decreased with increasing number of analyzed SNPs. However, the expected proportions of SNP homozygotes in linkage equilibrium with the average allele frequency (green dots in Figure 2) were much lower than the observed proportions of SNP homozygotes. This may be because the analyzed SNPs within the regions are in linkage disequilibrium. We measured the levels of linkage disequilibrium in the analyzed regions by a multi-locus linkage disequilibrium parameter, ε.[35] The average values of ε for each number of SNPs were

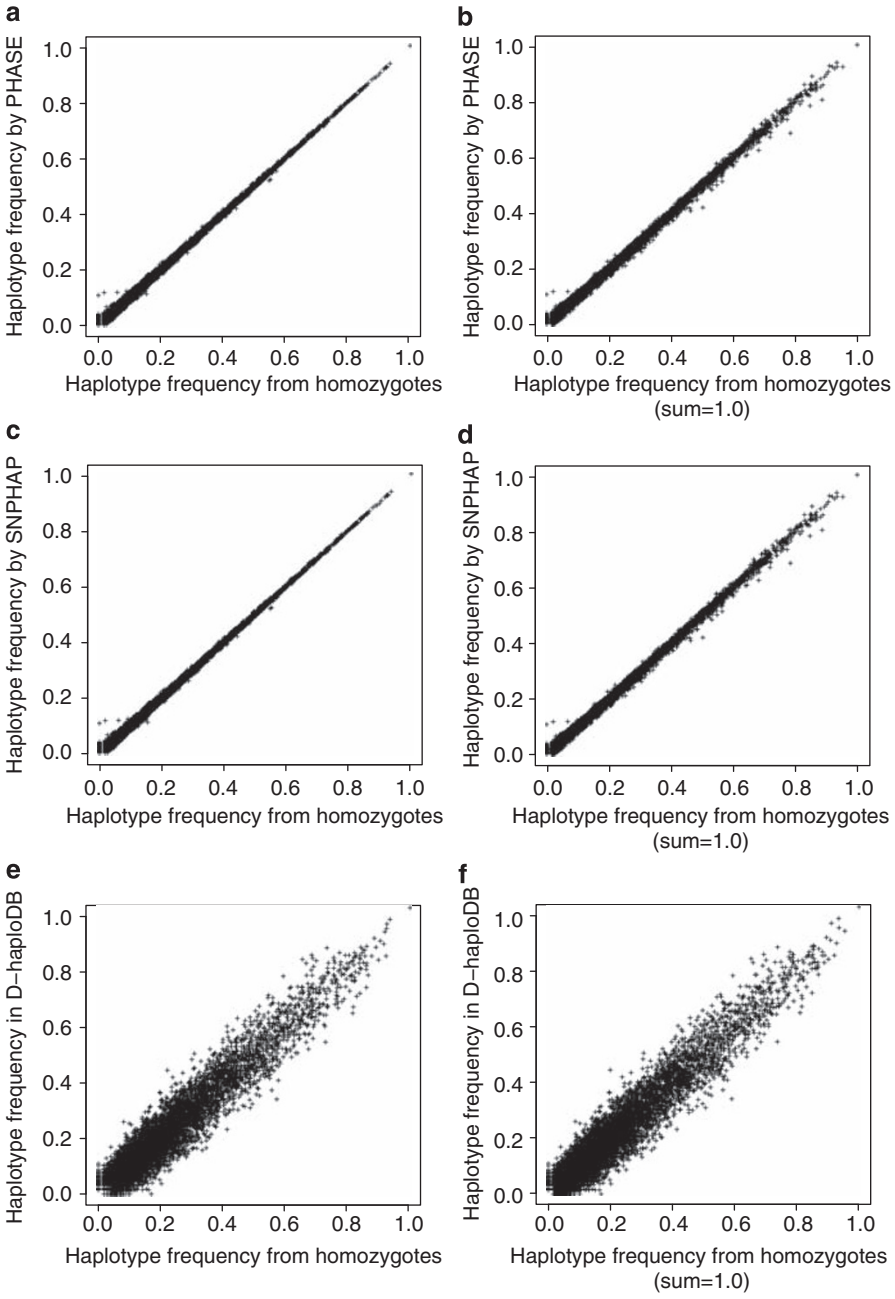

Figure 1 Comparison of haplotype frequencies from different approaches. (**a**–**f**) Haplotype frequencies obtained by two approaches are shown in scatter plots. *X* axis is haplotype frequency estimated from the proportion of homozygotes. (**a**–**d**) Comparison of haplotype frequencies with those estimated by two statistical methods, PHASE program (**a**, **b**) and SNPHAP program (**c** and **d**). Genomic regions having more than seven SNPs were not analyzed by those programs because of the long computation time they require. (**e**, **f**) Comparison of haplotype frequencies with those in the 74 CHMs from D-haploDB. (**b**, **d**, **f**) The haplotype frequencies estimated from the proportion of homozygotes were normalized so that the sum of frequencies of the observed haplotypes for each region equaled 1.0. The correlation coefficients of haplotype frequencies for each plot are given below; (**a**) 0.9986, (**b**) 0.9983, (**c**) 0.9985, (**d**) 0.9981, (**e**) 0.9691 and (**f**) 0.9688.

compared with those in the test analysis with the definitive haplotypes. The average ε values were always larger than those in the test analysis for various numbers of SNPs (see Supplementary Table 2). This may be due to a higher level of linkage disequilibrium between the analyzed SNPs that were more densely distributed in the regions compared with the test analysis.

If there are no unobserved haplotypes or unobserved haplotypes are negligible in terms of frequency, the sum of the estimated frequencies of the haplotypes would be 1.0 or very close to 1.0. However, as the number of the analyzed SNPs in the region increases, the number of actual haplotypes increases and some haplotypes in the sample may not be detected as SNP homozygotes. Therefore, we examined how much the sum of the estimated frequencies of the haplotypes deviated from 1.0. The sum of the haplotype frequencies was nearly 1.0 for the regions in which the numbers of analyzed SNPs were less than 20 (Table 3). However, as the number of analyzed SNPs increased above

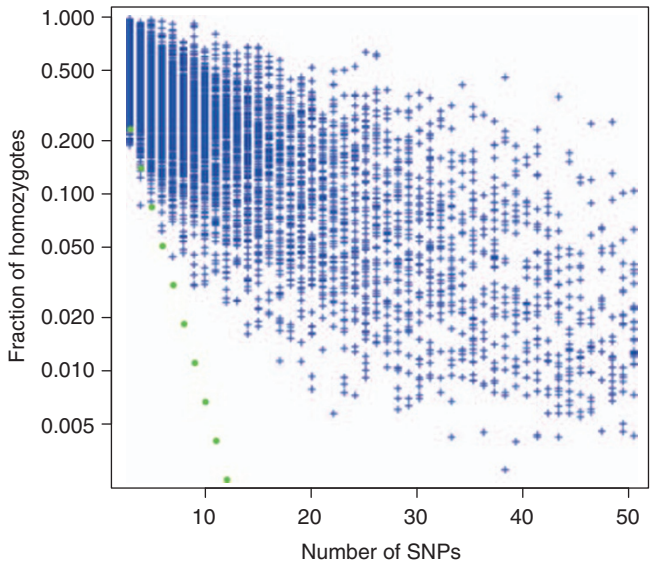

Figure 2 Proportion of SNP homozygotes in the 3397 individuals. The observed proportion of SNP homozygotes (+) for each analyzed region was plotted (*y* axis in log-scale) according to the number of analyzed SNPs (*x* axis). Green dots indicate the expected proportions of SNP homozygotes in linkage equilibrium, which were calculated with the average minor allele frequency (0.27).

Table 3 Total frequency of observed haplotypes from SNP homozygotes

No. of SNPs[a]	*No. of regions*	*Average length (bp)*	*Total frequency of haplotypes*[b]
3–9	7184	39 196	0.999
10–19	2151	90 372	0.954
20–29	739	138 529	0.847
30–39	349	188 832	0.742
40–49	223	185 804	0.619

Abbreviation: SNP, single-nucleotide polymorphism.
[a]The analyzed genomic regions with less than 50 analyzed SNPs were divided into five classes according to the number of SNPs.
[b]We calculated total frequency of the observed haplotypes for each genomic region, and the average of the total haplotype frequency is shown.

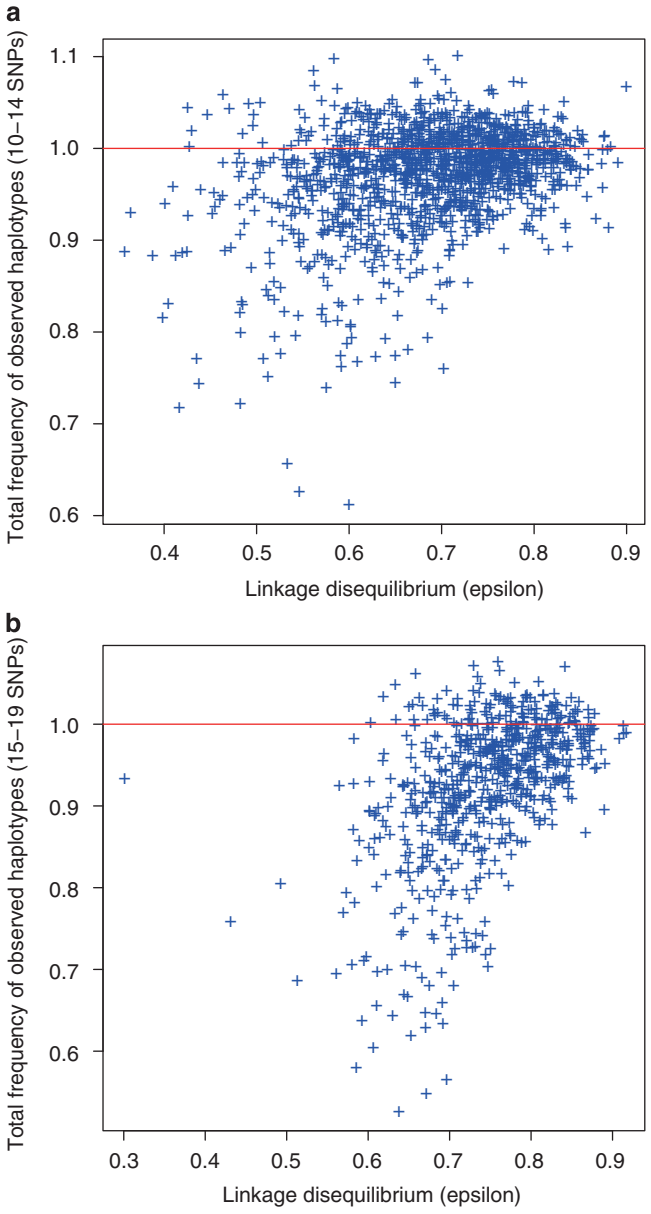

Figure 3 Relationship between the total frequency of the observed haplotypes and linkage disequilibrium. For the genomic regions with 10–19 analyzed SNPs, the total frequency of the observed haplotypes in SNP homozygotes for each region was calculated and was plotted according to the level of linkage disequilibrium (*x* axis) measured by epsilon (Nothnagel *et al.*[35]). The horizontal red line shows that the total frequency is 1.0. (**a**) Genomic regions with 10–14 analyzed SNPs. The correlation coefficient was 0.327 ($P<2.2\times10^{-16}$). (**b**) Genomic regions with 15–19 analyzed SNPs. The correlation coefficient was 0.497 ($P<2.2\times10^{-16}$).

20, the average sum of the haplotype frequencies became much smaller than 1.0 (for example, the average was 0.742 for the regions with 30–39 SNPs, Table 3). This suggests that the sum of frequencies of unobserved haplotypes becomes substantial when the number of SNPs is larger than 30 ($\sim$150 kb in this analysis). The total frequency of the observed haplotypes was positively correlated with the level of linkage disequilibrium for the genomic regions that had 10–19 analyzed SNPs (Figure 3). This may be because the genomic regions in higher linkage disequilibrium have fewer numbers of haplotypes that are less likely to be missed in detection as homozygotes.

DISCUSSION

Although our approach to determining haplotypes based on SNP homozygotes uses only homozygous individuals in a sample, the haplotype frequencies estimated from the proportion of homozygotes were quite similar to those obtained by conventional statistical approaches (PHASE and SNPHAP).[9] As mentioned above, some low-frequency haplotypes identified by statistical approaches may not be real. Two advantages of our approach, which is based on genotype data from thousands of individuals, are that it can detect real haplotypes without ambiguity and can estimate their frequencies.

The estimated haplotype frequencies were also similar to the frequencies of the 'definitive haplotypes' in 74 CHMs,[27,34] although the sample used in the previous studies was different from ours. Although the haplotypes determined from the genotype data of the 74 CHMs in D-haploDB are also real, the chromosomes in the CHMs might have accumulated mutations during their abnormal development. Our study identified 2808 additional haplotypes for the 1955

genomic regions that were not detected in the 74 CHMs. This was expected because their sample is different from ours and also much smaller than ours. Knowing haplotypes without ambiguity should improve haplotype estimation by statistical approaches.[36] Therefore, the haplotypes determined by our approach appear to be useful for haplotype inference with new data.

Our method, when applied to genotype data from about 3000 individuals, can detect haplotypes with frequencies as low as 0.03. Because the size of the data for genome-wide SNPs is increasing as more people are genotyped, our approach would be useful for identifying many real haplotypes including low-frequency haplotypes in a population. For example, in a sample of 5000 individuals, the probability of finding a haplotype whose frequency is 0.03 in a homozygous individual would be 98.9%.

To use our approach, the number and type of SNPs to be analyzed in a genomic region should be chosen with care because the efficiency of this approach depends on the nature of the SNP genotype data. As the number of analyzed SNPs increases, our method can obtain haplotypes on a finer scale. However, the proportion of homozygotes in a sample decreases with increasing number of analyzed SNPs, even though the number of existing haplotypes in the population is large. Therefore, our approach may fail to detect a large number of low-frequency haplotypes. When we applied this approach to a large number of genomic regions using genotype data from the Japanese population, the sum of frequencies of unobserved haplotypes appeared to be negligible when the analyzed region was less than 100 kb. It is also important to discard SNPs whose frequencies are low (for example, minor allele frequency <0.05) before conducting an analysis in order to avoid generating many low-frequency haplotypes that do not need to be distinguished in association studies. The criteria for designing an analysis may be modified depending on the level of diversity and history of the target population. The criteria can also be modified according to the different levels of linkage disequilibrium among local genomic regions, even in an analysis focusing on one population.

Another reason for limiting the length of the target region is that longer target regions are more likely to have a recent recombination that created new haplotypes. These new and younger haplotypes may exist only in limited local areas in the population, or their distributions among local geographic regions may not be uniform. In such a situation, combinations of haplotypes in individuals in the entire population may depart from Hardy–Weinberg equilibrium. On the other hand, the genotype frequencies at each SNP could be in Hardy–Weinberg equilibrium when there is little difference in SNP allele frequency among local geographic regions.

Although additional haplotypes can be identified by including individuals with only one heterozygous SNP, we did not use additional haplotypes for the following reasons. First, the contribution of these additional haplotypes would be very low in terms of frequency when thousands of individuals are analyzed. In fact, we detected additional haplotypes by allowing only one heterozygous SNP, and examined total frequencies of these haplotypes using the result obtained by the SNPHAP program. We found that the total frequencies of these haplotypes were very low (0.014 on average, shown in Supplementary Figure 5). Thus, a list of haplotypes detected in homozygotes would be enough to create a useful haplotype catalog for a GWAS, if a sufficient number of individuals were included in the analysis. Second, additional haplotypes that are detected on only one chromosome should be treated with caution, because some of them may be the result of a genotyping error in a homozygous individual and not really exist. In contrast, determining haplotypes by using only homozygotes would be robust against such a problem. Third, including such a limited proportion of heterozygotes would complicate estimation of haplotype frequencies.

Our approach can determine haplotypes without ambiguity, can estimate haplotype frequencies and can act as a catalog of haplotypes for the population. Another advantage of determining haplotypes by the proportions of homozygotes is that it does not take much computation time even if there is a number of SNPs in the genomic region of interest. The catalog of real haplotypes with their estimated frequencies will be useful for identifying causative polymorphisms for a trait, which are linked to the most associated SNPs in a GWAS. Furthermore, the catalog of haplotypes will be useful for haplotype-based GWAS[37,38] and detection of shared haplotypes that contain multiple variants[39,40] that affect the trait.

ACKNOWLEDGEMENTS

We thank Ryo Yamada, Kazuharu Misawa, Yukinori Okada and Koichiro Higasa for helpful discussion and comments on this study, and Yoshiyuki Yukawa for his technical assistance. We also thank all the members in the BioBank Japan Project for their effort in organizing the project and collecting samples. This study was supported by the Ministry of Education, Culture, Sports, Science and Technology.

1 Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. *et al.* The structure of haplotype blocks in the human genome. *Science (New York, NY)* **296,** 2225–2229 (2002).
2 Tsunoda, T., Lathrop, G. M., Sekine, A., Yamada, R., Takahashi, A., Ohnishi, Y. *et al.* Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. *Hum. Mol. Genet.* **13,** 1623–1632 (2004).
3 Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29,** 233–237 (2001).
4 Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
5 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* The sequence of the human genome. *Science (New York, NY)* **291,** 1304–1351 (2001).
6 Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P. *et al.* The diploid genome sequence of an individual human. *PLoS Biology* **5,** e254 (2007).
7 Clark, A. G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7,** 111–122 (1990).
8 Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12,** 921–927 (1995).
9 Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978–989 (2001).
10 Niu, T. Algorithms for inferring haplotypes. *Genet. Epidemiol.* **27,** 334–347 (2004).
11 Niu, T., Qin, Z. S., Xu, X. & Liu, J. S. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70,** 157–169 (2002).
12 Ito, T., Chiku, S., Inoue, E., Tomita, M., Morisaki, T., Morisaki, H. *et al.* Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am. J. Hum. Genet.* **72,** 384–398 (2003).
13 Lin, S., Cutler, D. J., Zwick, M. E. & Chakravarti, A. Haplotype inference in random population samples. *Am. J. Hum. Genet.* **71,** 1129–1137 (2002).
14 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84,** 210–223 (2009).
15 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
16 Mano, S., Yasuda, N., Katoh, T., Tounai, K., Inoko, H., Imanishi, T. *et al.* Notes on the maximum likelihood estimation of haplotype frequencies. *Ann. Hum. Genet.* **68** (Part 3), 257–264 (2004).
17 Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513–516 (2000).
18 Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9,** 677–679 (1999).
19 Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. & Tanaka, T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome. Single-nucleotide polymorphism. *J. Hum. Genet.* **47,** 605–610 (2002).

20 Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. & Nakamura, Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res* **30,** 158–162 (2002).
21 Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32,** 650–654 (2002).
22 Suzuki, A., Yamada, R., Chang, X., Tokuhiro, S., Sawada, T., Suzuki, M. *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **34,** 395–402 (2003).
23 Altshuler, D. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
24 Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
25 Consortium TWTCC. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **447,** 661–678 (2007).
26 Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78,** 437–450 (2006).
27 Kukita, Y., Miyatake, K., Stokowski, R., Hinds, D., Higasa, K., Wake, N. *et al.* Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. *Genome Res.* **15,** 1511–1518 (2005).
28 McQuillan, R., Leutenegger, A L., Abdel-Rahman, R., Franklin, C S., Pericic, M., Barac-Lauc, L. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83,** 359–372 (2008).
29 Kong, A., Masson, G., Frigge, M L., Gylfason, A., Zusmanovich, P., Thorleifsson, G. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40,** 1068–1075 (2008).
30 Seelow, D., Schuelke, M., Hildebrandt, F. & Nurnberg, P. HomozygosityMapper—an interactive approach to homozygosity mapping. *Nucleic Acids Res* **37** (Web Server issue), W593–W599 (2009).
31 Lencz, T., Lambert, C., DeRosse, P., Burdick, K E., Morgan, T V., Kane, J M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl Acad. Sci. USA* **104,** 19942–19947 (2007).
32 Thiadens, A A., den Hollander, A I., Roosing, S., Nabuurs, S B., Zekveld-Vroon, R C., Collin, R W. *et al.* Homozygosity mapping reveals PDE6C mutations in patients with early-onset cone photoreceptor disorders. *Am. J. Hum. Genet.* **85,** 240–247 (2009).
33 Nakamura, Y. The BioBank Japan Project. *Clin. Adv. Hematol. Oncol.* **5,** 696–697 (2007).
34 Higasa, K., Miyatake, K., Kukita, Y., Tahira, T. & Hayashi, K. D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples. *Nucleic Acids Res.* **35** (Database issue), D685–D689 (2007).
35 Nothnagel, M., Furst, R. & Rohde, K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum. Hered.* **54,** 186–198 (2002).
36 Higasa, K., Kukita, Y., Kato, K., Wake, N., Tahira, T. & Hayashi, K. Evaluation of haplotype inference using definitive haplotype data obtained from complete hydatidiform moles, and its significance for the analyses of positively selected regions. *PLoS Genet.* **5,** e1000468 (2009).
37 Misawa, K., Fujii, S., Yamazaki, T., Takahashi, A., Takasaki, J., Yanagisawa, M. *et al.* New correction algorithms for multiple comparisons in case-control multilocus association studies based on haplotypes and diplotype configurations. *J. Hum. Genet.* **53,** 789–801 (2008).
38 Tregouet, D A., Konig, I R., Erdmann, J., Munteanu, A., Braund, P S., Hall, A S. *et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* **41,** 283–285 (2009).
39 Cohen, J C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G L., Grundy, S M. *et al.* Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl Acad. Sci.USA* **103,** 1810–1815 (2006).
40 Pritchard, J K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69,** 124–137 (2001).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)