European Journal of Human Genetics (2007) 15, 302–312
© 2007 Nature Publishing Group All rights reserved 1018-4813/07 $30.00
www.nature.com/ejhg

ARTICLE

Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations

Sheng Gu[1], Andrew J Pakstis[1], Hui Li[1], William C Speed[1], Judith R Kidd[1] and Kenneth K Kidd*,[1]

[1]*Department of Genetics, Yale University School of Medicine, New Haven, CT, USA*

The initial belief that haplotype block boundaries and haplotypes were largely shared across populations was a foundation for constructing a haplotype map of the human genome using common SNP markers. The HapMap data document the generality of a block-like pattern of linkage disequilibrium (LD) with regions of low and high haplotype diversity but differences among the populations. Studies of many additional populations demonstrate that LD patterns can be highly variable among populations both across and within geographic regions. Because of this variation, emphasis has shifted to the generalizability of tagSNPs, those SNPs that capture the bulk of variation in a region. We have examined the LD and tagSNP patterns based upon over 2000 individual samples in 38 populations and 134 SNPs in 10 genetically independent loci for a total of 517 kb with an average density of 1 SNP/5 kb. Four different 'block' definitions and the pairwise LD tagSNP selection algorithm have been applied. Our results not only confirm large variation in block partition among populations from different regions (agreeing with previous studies including the HapMap) but also show that significant variation can occur among populations within geographic regions. None of the block-defining algorithms produces a consistent pattern within or across all geographic groups. In contrast, tagSNP transferability is much greater than the similarity of LD patterns and, although not perfect, some generalizations of transferability are possible. The analyses show an asymmetric pattern of tagSNP transferability coinciding with the subsetting of variation attributed to the spread of modern humans around the world.

European Journal of Human Genetics (2007) **15,** 302–312. doi:10.1038/sj.ejhg.5201751; published online 3 January 2007

Keywords: haplotype; haplotype block; linkage disequilibrium; tagSNPs

Introduction

Haplotype and linkage disequilibrium (LD) analyses are widely used for mapping disease genes and understanding human population history.[1–4] Some haplotype studies have suggested that the human genome can be parsed objectively into haplotype blocks in which there is little evidence for historical recombination.[5–9] In recognition of a few common haplotypes capturing most of the genetic variation across regions encompassing multiple SNPs, Gabriel *et al*[7] proposed testing only the minimum number of SNPs sufficient to define the common haplotypes, haplotype tag SNPs (htSNPs or tagSNPs). Defining the genomic regions of high LD, that is, haplotype 'blocks', was an initial focus and continued to be an interest of the International HapMap project.[8,9]

The effort to construct a haplotype map or maps of the human genome based on limited population samples has

*Correspondence: Dr KK Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, CT 06520-8005, USA.
Tel: +1 203 785 2654; Fax: +1 203 785 6568;
E-mail: kenneth.kidd@yale.edu

Received 21 July 2006; revised 26 September 2006; accepted 25 October 2006; published online 3 January 2007

an underlying assumption of extensive sharing of patterns of LD such that tagSNPs will show good 'transferability' (how well tags selected in one or more analysis panels will transfer to disease studies performed in these or other populations[9]) among populations, especially populations within the same geographic regions. Despite ongoing controversy over some issues,[10–12] the original belief that haplotype blocks represent a fundamental aspect of the human genome appears to be an oversimplified view of genome organization.

Few studies have examined LD in a global set of populations. One study used the Human Genome Diversity Project-CEPH panel to examine variation in LD at one region on chromosome 22. It concluded that a wide communality in LD patterns exists in human populations from different continental regions despite differences in their demographic histories.[13] However, this conclusion was based on only 12 SNPs across a 1.78 Mb region. In many populations, there was little LD, which resulted in similar tagSNP sets. Another much more extensive study of the portability of tagSNPs across global populations at higher marker density focused just on tagSNP transferability.[14] It concluded that tagSNPs from one population are highly informative in other populations within each continental group. However, other studies have reported significant variation among populations in block structure and tagSNPs. Sawyer *et al*[15] studied three loci in 16 diverse populations with an emphasis on African and European populations. They found significant quantitative and qualitative variation in LD among populations both across and within geographic groups, and no group showed consistency in patterns of LD for all three loci under study. Liu *et al*[16] reached a similar conclusion with respect to tagSNPs.

The HapMap data do not address directly the question of how tags selected in one or more of the populations studied will transfer to disease studies performed in these or other populations, and thus the general applicability of the HapMap data needs to be confirmed in samples from several local populations.[8,9] We are examining a set of populations comparable to the HGDP-CEPH panel used by González-Neira *et al*;[13,14] indeed, many of the population samples in that panel originated in our lab. However, our sample sizes are, on average, twice as large as in the HGDP-CEPH panel. Using data that represent an even more diverse sampling of LD patterns among populations than initially planned for extensions of the HapMap effort, we consider consistency of blocks and tagSNPs among populations and the similarity of different methods in defining blocks. Our data on 38 populations widely distributed around the world (Figure 1) illustrate considerable diversity in the patterns of LD suggesting that HapMap data do not generalize in this respect. However, the transferability of tagSNPs is, in general, quite high, in agreement with González-Neira *et al*.[14]

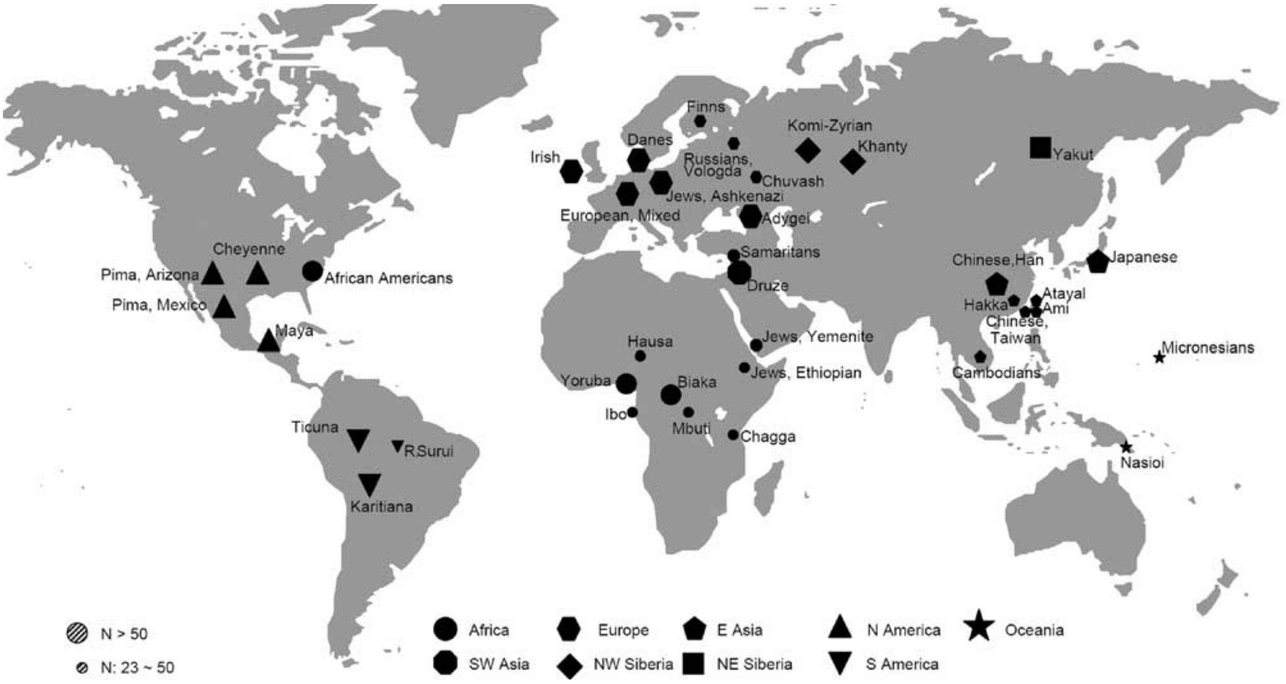

Figure 1 The graphical distribution of the 38 population samples used in this study. Populations are categorized into nine different geographic groups according to their ancestry. The large symbols stand for a sample size of over 50, whereas the small symbols represent a sample size from 23 to 50.

Subjects and methods

Population samples

A total of approximately 2000 individuals from a global sample of 38 human populations (Figure 1) were typed in this study. According to their ancestry, all populations are categorized into nine geographic groups: Africa, Southwest Asia, Europe, Northwest Siberia, East Asia, Oceania, Northeast Siberia, North America and South America.[17] The sample sizes of populations range from 23 (Nasioi) to 118 (Irish), with most having around 50 individuals (Supplementary Table 1). Detailed information on the individual populations and samples is in ALFRED (the ALlele FREquency Database).[18,19]

Genomic regions and SNPs

Ten unlinked genomic regions on seven chromosomes encompassing a total length of 338 kb are considered in this study: CCR5, CD4, COMT, DRD2-NCAM1, HOXA, HOXB, PAH, RET-D10S94, SORCS3 and THRAP, with a total of 134 SNPs (including one insertion-deletion site). We selected 10 loci that had regions of approximately 40–50 kb with an average SNP density of approximately 1 SNP per 5 kb with complete typing on 38 population samples. Some SNPs used in this study were also included in the HapMap project, whereas others have only been genotyped in our populations. The relative positions and densities of markers in both the HapMap project and our data are shown in parallel (Figure 2). These are genes that have been foci for diverse studies to define haplotype frequencies with no selection for pattern of LD. These regions were chosen for this analysis because we had achieved a density of ~1 informative SNP per 5 kb across 40–50 kb, as part of diverse ongoing studies. No selection for pattern of LD was involved. SNPs were selected for analysis because they had heterozygosity greater than ~10% in both African and non-African populations. There was no selection for a more specific allele frequency pattern based on available information. These markers provided more uniform spacing than alternatives. Some additional SNPs were typed, but had uniformly low heterozygosity or were over 10 kb away from the analyzed segment. Overlapping data at RET-D10S94 and DRD2-NCAM1 were in Sawyer *et al.*[15] The detailed information on selected genomic regions and SNPs is listed in Supplementary Table 2. Taqman® was the main genotyping method with a small subset of SNPs genotyped with Fluorescence Polarization and gel-based restriction fragment length polymorphism methods.

The phase-unknown marker typing data used in this study can be downloaded from http://info.med.yale.edu/genetics/kkidd/contents.html. Allele frequencies and the sample sizes of all sites can be found in ALFRED under the site UIDs (Supplementary Table 2). For each marker, the typing was complete for at least 90% of the individuals in all populations. For each site, the average heterozygosity and Fst value[20] over 38 population samples are listed in Supplementary Table 3.

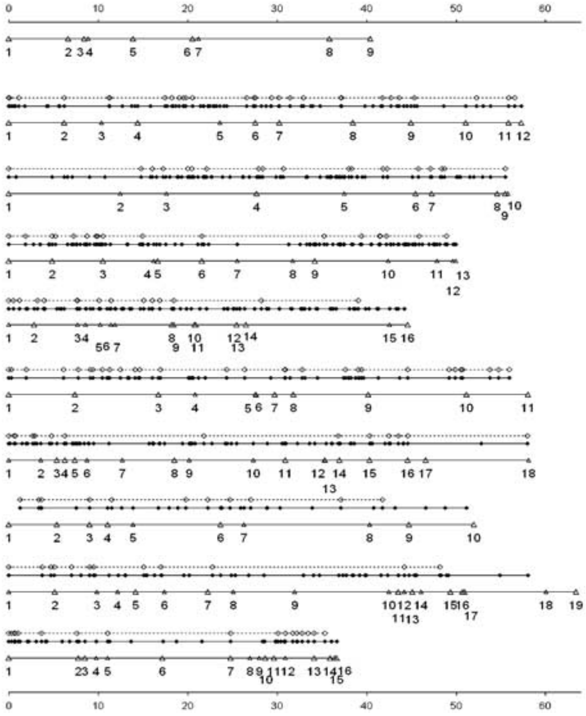

Figure 2 The 10 loci used in this study. All SNPs are numbered in the same order as in Supplementary Table 2. For each locus, the bottom solid lines with hollow triangles show the marker density of our own data, whereas the middle solid lines show the marker density of corresponding regions from the HapMap project. Big triangle symbols across the bottom solid lines stand for the tagSNPs selected for our European American population through the pairwise LD method, and the diamond symbols across the top dotted lines stand for the tagSNPs selected for CEU population from the HapMap project.

Statistics and analyses

The haplotype block patterns in all 38 populations are presented using the HAPLOT application.[21] Four different block definitions are implemented in HAPLOT and used in this study. The definitions based on the LD measure D' and its confidence interval (GAB), the four-gamete test (4GA), and the solid spine of LD (SPI) are available in Haploview v3.2 and thus are called by HAPLOT directly for plotting.[7,22,23] The fourth definition, named Kidd r^2(KR2), is a block partition algorithm based on the LD measure r^2.[21] The statistical comparison of the block partitions uses the Liu *et al*[16] similarity measure. This similarity measure is an improved version of Fisher's exact test that considers not only the number of partitions but also the boundaries of partitions. It tests the null hypothesis that there is no correspondence between two block partitions.

Initially three different tagSNP selection algorithms were applied in this study: the haplotype diversity method (HD), the haplotype entropy method (HE) and the pairwise LD method (LD).[24–27] For the first two methods, by default the common haplotype threshold is set at 0.05 and the fraction of haplotype diversity/entropy explained by tagSNPs is set at 0.8. For the pairwise LD method, by default the r^2 threshold is set at 0.8 and the LOD threshold for multi-marker tests is set at 3.0. In HAPLOT, we developed and implemented algorithms to solve the ambiguity problem associated with tagSNP selection.[21] Pilot studies suggest that with the same parameter settings, HD and HE methods are much more sensitive to the marker density than is the LD method (Supplementary Figure 1). The increase of SNP numbers adds more complexity and variation to the haplotype composition, and the methodology based upon the preservation of haplotype variation, such as the HD algorithm, is sensitive to the number of component SNPs (except fixed SNPs or SNPs with extremely low heterozygosity). The LD method responds more to the strength of LD between SNPs than to the number of component SNPs, and thus it is less sensitive to the increased marker density. The hypersensitivity to marker density makes the HD and HE methods less efficient and reliable compared with the LD method. Thus, for our studies, we focused only on the LD method for the tagSNP variation and transferability analyses.

We note that the relative savings from using tagSNPs is a function of the initial density of SNPs from which tagSNPs are selected. In Figure 2, the LD tagSNP selection algorithm has been applied to the CEU population data from the HapMap project and to data on our European American population. The CEU population has a marker density higher than 1 SNP/kb in all loci (except CCR5, where there are no corresponding SNPs from the HapMap project), whereas the European American population has a marker density around 1 SNP/5 kb. The average saving of SNPs by using tagSNPs from the CEU populations is much higher than that from the European American population, because with higher marker density many SNPs are physically close and in strong LD with each other, which causes each selected tagSNP to be associated with more unselected SNPs. However, the savings is not strictly proportional as marker spacing is also relevant.

TagSNP transferability is represented by the percentage of captured common variants (with a frequency of $\geq 5\%$) of Pop2 by the tagSNPs from Pop1.[28] If an untagged SNP has an r^2 value with any tagSNP above the threshold (we use 0.64 and 0.8 in our study), we consider it captured by the tagSNP set. For each pair of populations being compared, the tagSNP transferability values are unidirectional and asymmetrical. For example, tagSNPs from Pop1 (a larger collection of tagSNPs) could preserve 100% variation of Pop2 whereas tagSNPs from Pop2 (a smaller collection of tagSNPs) could only preserve a small fraction of variation of Pop1.

Results

Haplotype block variation

Haplotype block structure is highly variable among populations both within and across geographic regions (Figure 3; Supplementary Figures 2–10). For each locus, four different block partition algorithms have been applied. Figure 3 uses the PAH locus as a typical example to illustrate high block variation among populations no matter what partition algorithm has been used. The HAPLOT plots show all identified haplotype blocks in all 38 populations, and the corresponding GOLD[29] plots show the significance levels of the statistical comparisons (Liu *et al*[16] similarity measure) of block partitions for each pair of populations. None of the GAB, 4GA and SPI partition algorithms suggests any block patterns associated with the categorization of geographic groups. Thus, the variation within groups can be as high as that across groups. The KR2 partition method based upon r^2 values generally shows greater consistency of block patterns among populations in the same geographic regions than the other methods, but all methods show variation among loci. In general, greater 'consistency' of LD structure among populations of similar genetic background than between populations from different backgrounds (eg, geographic regions) is locus-dependent. For example, using the KR2 partition method, African populations show very similar LD patterns at the PAH locus (Figure 3), but the opposite is seen at the CCR5 region (Supplementary Figure 2). The change of parameter thresholds in the KR2 algorithm could change number and length of blocks in individual populations, but not the general pattern across different geographic regions. For example, if we enforce a stricter rule (higher r^2) in block partition, some blocks may disappear or become shorter because of failing to meet the increased r^2 threshold. However, this change is global. Populations of the same geographic groups still tend to be more consistent in block patterns than across geographic groups (data not shown).

Similar variation in LD patterns within and among geographic regions can be seen at the other loci studied (Supplementary Figures 2–10). In summary, the idea of haplotype blocks consistent across populations, even within geographic regions, is an over-simplified view of patterns of LD. Our results provide an empirical validation of the simulations of the haplotype block model by Wall and Pritchard.[30]

TagSNP transferability and variation

Of greater relevance to biomedical research is tagSNP transferability. Our results for tagSNP transferability show the generalizability of the portability of tagSNPs in spite of the considerable heterogeneity in the pattern of LD among populations. González-Neira *et al*[14] not only emphasized the average level of transferability but also showed considerable variation among individual SNPs. Here, we look at the tagSNP transferability of all loci in three

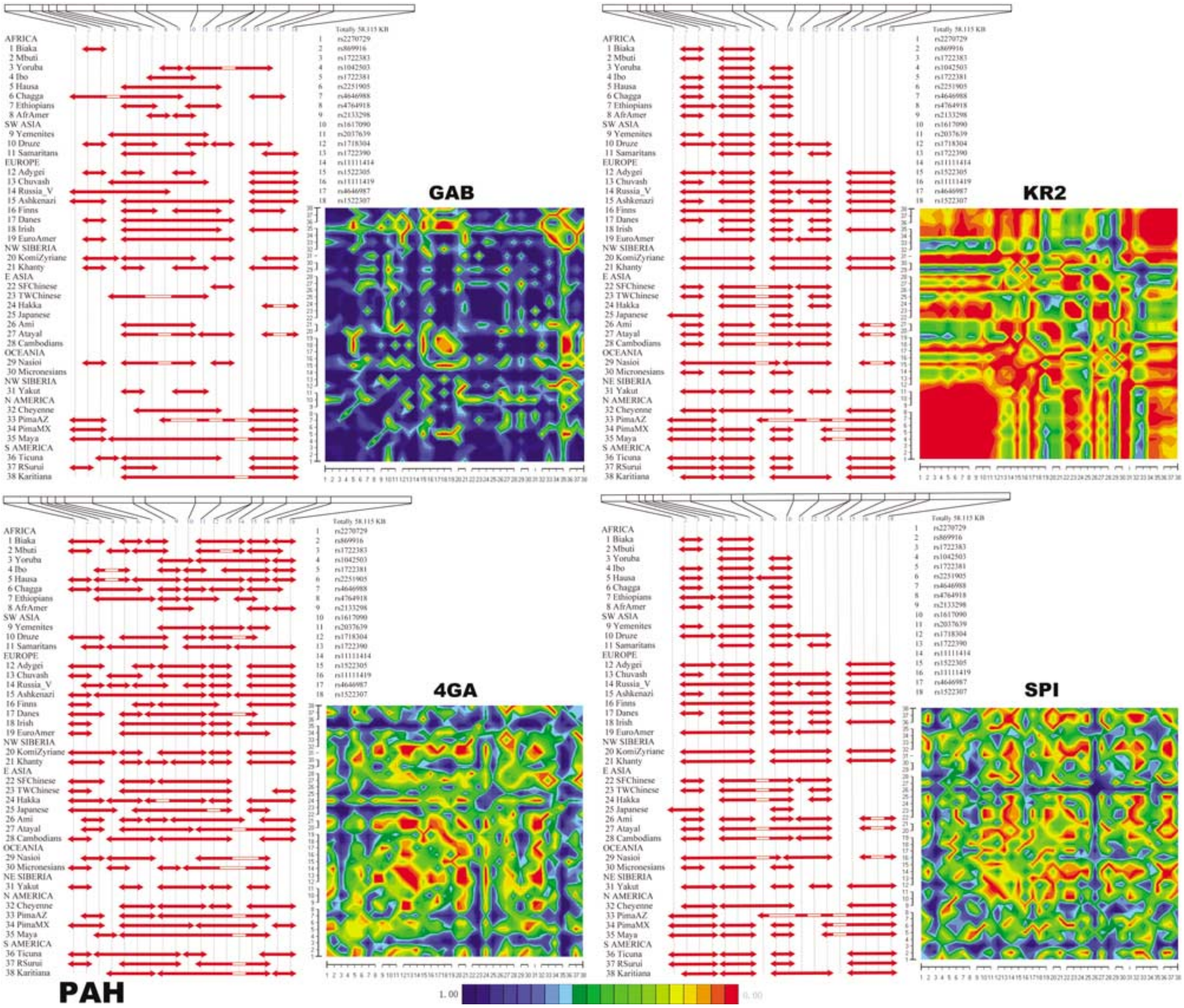

Figure 3 The haplotype block patterns and the pairwise similarity tests of block patterns in all 38 populations at the PAH locus. The HAPLOT plots show schematic maps of SNPs on the top, the population names on the left and the site names on the right. Blocks are represented by the double-arrowed segments. The SNPs inside blocks that fail Hardy–Weinberg tests or are below 0.05 heterozygosity level are indicated by the uncolored part of the segments. The corresponding GOLD plots show the pairwise statistical significance of the Liu *et al*[16] similarity tests with numbers on the axes corresponding to the population numbers. The color scheme of GOLD plots is based upon the *P*-values, with the bright red representing *P*-values below 0.05 (most significant), rejecting the null hypothesis of no similarity.

geographic groups and find variation by both geographic and genomic region. We chose our Yoruba, European American and Japanese samples as our own reference populations to compare with other populations of the same geographic regions. Despite the fact that the YRI, CEU and JPT populations in the HapMap project are different from our Yoruba, European American and Japanese samples, we assume that our Yoruba, European American and Japanese samples are closest to those HapMap reference populations. When the tagSNP selection and evaluation threshold for r^2 is set at 0.8, applying tagSNPs from Yoruba to other African populations captures an average of 95.1% of the common variants (ranges from 72.7 to 100% among genomic regions); applying tagSNPs from European American to other European populations captures an average of 94.9% of the common variants (ranges from 70.0 to 100%); applying tagSNPs from Japanese to other east Asian populations captures an average of 91.7% of the common variants (ranges from 55.6 to 100%) (Figure 4).

The pairwise portability of tagSNPs for all populations has been examined for all loci at two different r^2 thresholds: 0.64 and 0.80 (Figure 5). Each threshold has been applied to both the selection of tagSNPs and the detection

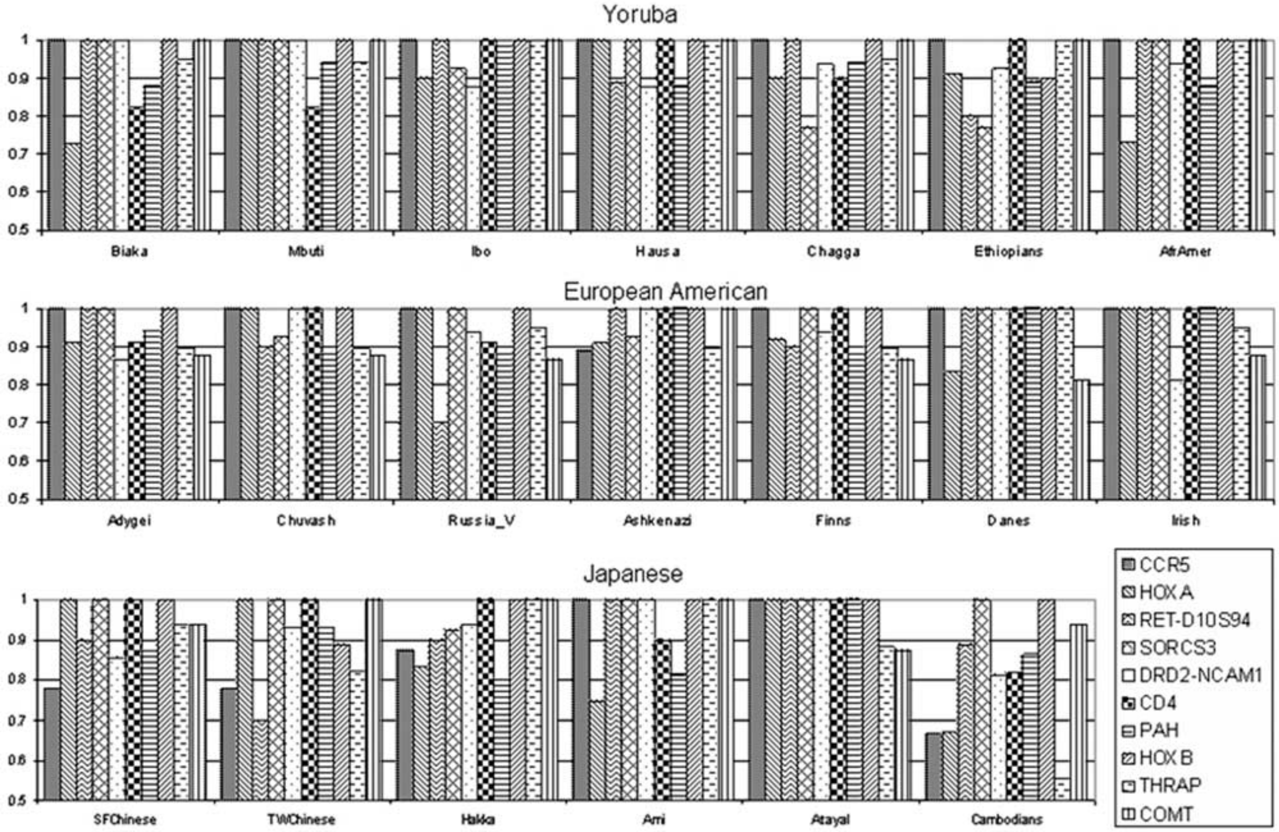

Figure 4 The percentages of common variants captured at $r^2 = 0.80$ by applying tagSNPs from designated reference populations to populations of the same groups. From the top to the bottom, three geographic groups have been studied: African, European and east Asian. The Yoruba, European American and Japanese populations have been used as reference populations of their own groups.

of common variants. For example, a tagSNP panel from the source population is first selected at r^2 level of 0.8, and then that panel has been applied to all other target populations with r^2 level of 0.8 for 'capture' of SNPs. A typical observation from Figure 5 is the asymmetry of the tagSNP coverage (the percentage of common variants captured by a source tagSNP set) between a pair of populations. For example, in CCR5 with either 0.64 or 0.80 selected as r^2 thresholds, the tagSNPs from Yoruba have a 100% coverage of Ami, whereas the tagSNPs from Ami capture only 44.4% of common variants in Yoruba. A general pattern shows that along a path from Africa to southwest Asia and then to Europe, east Asia, or the Americas, the tagSNPs selected from earlier populations along that path tend to provide a good coverage of later populations, but not vice versa. This pattern is consistent with the pathway predicted by the model of expansions of modern humans out of Africa and spreading across the globe. The farther apart geographically two populations are from each other the more asymmetric the tagSNP coverage.

Applying $r^2 = 0.80$ sometimes provides slightly better tagSNP portability among populations than applying $r^2 = 0.64$ (such as the COMT locus). Other times applying $r^2 = 0.64$ provides slightly better portability (such as east Asian and European populations at the HOXB cluster) (Figure 5). However, in general, two different r^2 thresholds provide similar levels of tagSNP portability. The major difference between these two different threshold levels is the number of selected tagSNPs. The 0.80 threshold enforces a stricter rule in determining the LD between a tagSNP and an untagged SNP, thus it usually results in having more SNPs in the tagSNP panel. Because two different thresholds are used in Figure 5, and each involves both selection and application, the color scheme only represents the tagSNP transferability level at its specific threshold. We should note that though the transferability is roughly the same, on average, using $r^2 = 0.64$ and $r^2 = 0.80$, the actual variation captured at $r^2 = 0.64$ is less than at $r^2 = 0.80$. We applied tagSNP selection criteria at $r^2 = 0.64$ and $r^2 = 0.80$, but set tagSNP detection threshold at $r^2 = 0.80$. The tagSNPs selected at $r^2 = 0.64$, when applied to detect common variants at $r^2 = 0.80$, fail to capture an appreciable amount of variation that could be captured by a tagSNP panel selected at $r^2 = 0.80$, and thus show lower transferability among populations (Figure 6).

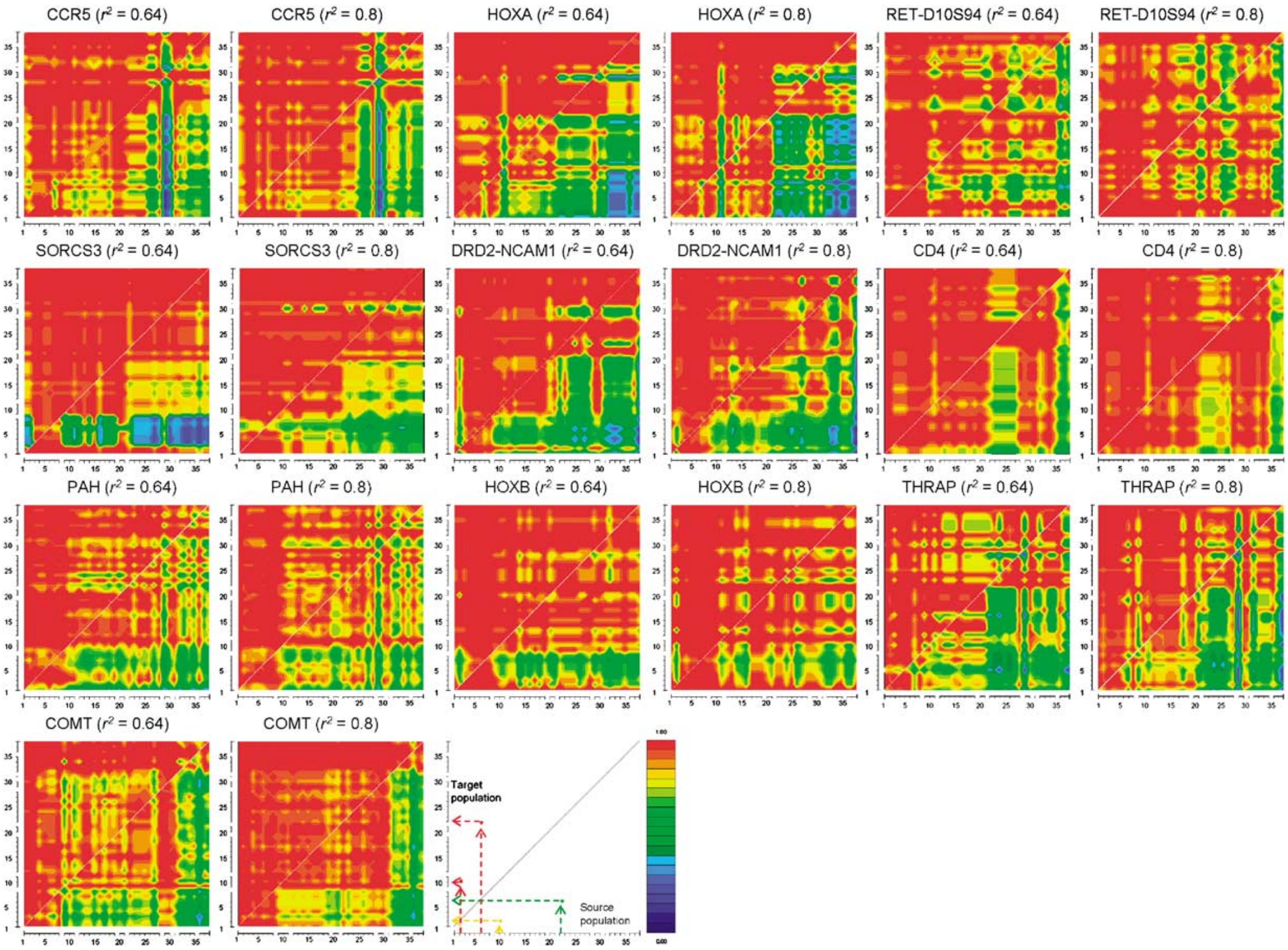

Figure 5 The percentages of common variants captured for $r^2 = 0.64$ and $r^2 = 0.80$ by applying tagSNPs selected at the corresponding r^2 from one population to another in all 38 populations. The vertical axis labels the target populations, whereas the horizontal axis labels the tagSNP source populations. The colored plot represents percentages of common variants captured in the target populations when tagSNPs from the source populations are applied. The asymmetry of tagSNP transferability is evident from the differences between the upper left and lower right halves.

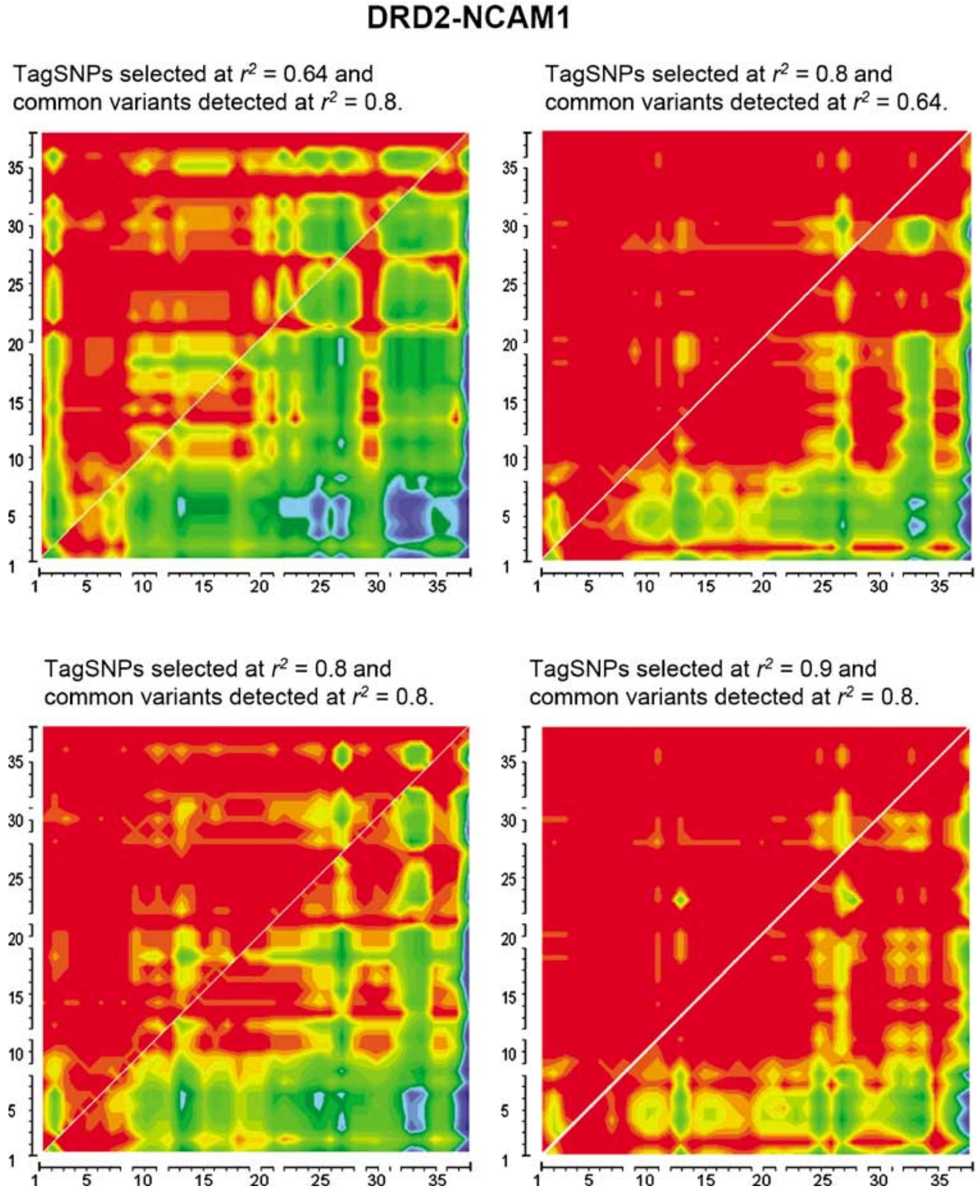

Figure 6 The percentages of common variants at DRD2-NCAM1 captured under four scenarios. One scenario with tagSNP selection at $r^2 = 0.64$ and detection in other populations at $r^2 = 0.80$, one with tagSNP selection at $r^2 = 0.8$ and detection at $r^2 = 0.64$, one with tagSNP selection and detection both using $r^2 = 0.80$, repeated from Figure 5, and one with tagSNP selection at $r^2 = 0.9$ and detection at $r^2 = 0.8$. The color scale is the same as in Figure 5. The additional tagSNPs selected at $r^2 = 0.90$ clearly provide best overall transferability when applied at $r^2 = 0.80$ detection, whereas the tagSNPs selected at $r^2 = 0.64$ have poorest transferability when applied at $r^2 = 0.80$ detection.

In Figure 5, we note variation among the loci in performance of tagSNP transferability and the amount of tagSNP variation. A typical example is the tagSNP portability among African populations for different loci. For RET-D10S94, CD4 and COMT, tagSNP portability is very high among different African populations. No matter what specific population is chosen as the reference population, its tagSNP panel is informative in other populations. However, for the SORCS3, DRD2-NCAM1 and HOXB loci, the selection of the reference population is critical. In SORCS3, Biaka's tagSNPs capture little variation in other African populations; at DRD2-NCAM1 and HOXB, Mbuti's tagSNPs fail to be representative of most of the other populations. Another example is the East Asian group at RET-D10S94 locus. TagSNPs from Ami, Atayal and Cambodians give poor coverage when applied to other East Asian populations.

Despite the slight variation described above, generally speaking, tagSNPs are highly informative in other populations within each continental group. In addition, African populations have the largest collection of tagSNPs and their tagSNPs are able to capture a large amount of variation of all other populations. TagSNPs defined in Europeans are often efficient for Asian (including southwest Asian, east Asian, northwest Siberian and northeast Siberian) populations, which agrees with the findings of González-Neira *et al.*[14]

Despite the above two arbitrarily selected r^2 thresholds, we could also apply any other thresholds. However, if the threshold is too low, unacceptable loss of information results. With higher r^2 threshold, the size of the tagSNP panel expands, but the asymmetric portability pattern remains similar (Figure 5 shows the pattern when applying the same threshold to the selection of tagSNPs and to the detection of common variants; Figure 6 shows the case when applying two different thresholds).

Discussion

The pattern of LD in a population is determined not only by the distribution of recombination events but also by both demographic factors determining the amount of random genetic drift and the purely chance aspects of whether a recombinant chromosome survives in the population. Moreover, at very low rates of recombination there is a stochastic aspect to whether or not a crossover will ever occur within the history of a population. Thus, it is not surprising that block patterns differ among populations. If a block is not a fundamental aspect of the genome, it is nonetheless a region of high LD within a specific population. We use *block* in that latter empiric sense.

The large amount of variation in haplotype block structure among global populations reflects the considerable haplotype variation among these populations seen at several loci in previous studies.[15,31] We infer from our data that random genetic drift is the major cause of LD patterns in different populations. One piece of supportive evidence is a comparison of block patterns between Africa and the Americas. African populations have relatively larger population size over a long history, and thus preserve more haplotypes than Native American populations, which are considered to have experienced bottleneck events and have much smaller population sizes historically.[32] Yet, the haplotypes outside of Africa tend to be a subset of those seen in Africa indicating little to no recent generation of new haplotypes by recombination. The general pattern shown by multiple loci is a progressive subsetting of haplotypes as one goes from Africa to southwest Asia then to Europe or to east Asia and separately to the Americas.[17,31] Thus, tagSNPs selected in an African population can discriminate among (identify) all haplotypes including the subsets that exist elsewhere. Similarly, the tagSNPs selected on any population will be likely to apply to populations further along in the subsetting, farther from Africa, but not in the reverse direction. Fewer and smaller multisite haplotype blocks are usually observed in African populations, whereas big haplotype blocks representing strong LD over large chromosome distance can often exist in Native American populations.

Our analyses of a global sample of 38 populations extends our previous findings that LD and tagSNP patterns differ among populations, even among populations of similar geographic origins.[15,16] Such differentiation is mainly due to the individual population demographic history and combined effects of genetic factors such as drift, mutation and recombination. Sampling error can also affect similarity, but at average sizes of 50 individuals is considered a minor factor.[33]

By studying the variation of haplotype block structure in all 10 representative loci using four different block definitions, we strove to disentangle the variation caused by the diversity of genetic background of different populations from that caused by the selection of any specific algorithm. As is shown (in Figure 3; Supplementary Figures 2–10), no single block definition shows consistency of block structure among populations of similar genetic background. Thus, our conclusion of high variation of block patterns is consistent with earlier observations from other researchers. However, by including a large global sampling of populations, we provide much more general evidence in support of the current shift of research focus from haplotype block structure to tagSNPs with a comprehensive and systematic study involving a representative number of SNPs and loci on different chromosomes. Nonetheless, the idea of a haplotype block, defined empirically as a region of high LD, although lacking generalizability across global populations, is considered useful for fine-scale mapping of complex traits if the study is restricted to localized populations.

TagSNP analyses lead to a slightly different conclusion: although a small amount of variation exists among populations within each continental group, in general the tagSNP portability is high. Thus, our results provide supportive evidence for the HapMap's endeavor of constructing haplotype maps based on a few reference populations. Our observation that tagSNPs are effective for both related and distant populations agrees with previous studies.[13,14] However, variation in tagSNP transferability does exist. The selection of reference populations together with the geographic categorization of populations is associated with the risk of losing information. For example, in CCR5, applying the Japanese tagSNP panel to Han Chinese (from San Francisco and Taiwan) will cause 22.8% loss of the variation. If applied to Cambodians, this loss becomes 33.3% (Figure 4). In THRAP, applying the Japanese tagSNPs to Cambodians will result in loss of 45% of the variation. In CCR5, the two Oceanian populations, Nasioi and Micronesians, can hardly be grouped, because their own tagSNP panels are not representative of any other populations (Figure 5). One safe way to guarantee a good capture of variation is to apply the tagSNP panel of the African reference population (Yoruba in this case) to all isolated populations such as Nasioi and Micronesians. However, as more and more new populations enter the study panel, this solution is not economically efficient.

Comparison of the transferability of tagSNPs selected at $r^2 = 0.64$ and $r^2 = 0.80$ (Figure 5) shows another general pattern: tagSNPs selected to meet the higher threshold generally perform better on other populations. This is expected because more SNPs are selected. Also, note that except in a few cases, tagSNPs do not fail to identify variation in other populations, they just identify less variation. This suggests a strategy that could be employed when studying a 'new' population. One can select tagSNPs from the closest HapMap population at a very stringent level, such as $r^2 = 0.90$, but expect that the level of transferability will be lower, say closer to $r^2 = 0.80$, but still useful unless the target population is very different from one of the HapMap populations. Although the result seems self-evident, at least qualitatively, we tried selecting at $r^2 = 0.90$ in each population and evaluating the % of SNPs covered at $r^2 = 0.80$ in target populations. Figure 6 illustrates this for DRD2-NCAM1, one of the regions showing the lowest transferability at $r^2 = 0.80$ as measured by the blocks of blue and green in Figure 5. The expected result is clearly seen.

It might be true that new haplotype maps in populations other than the four HapMap reference populations are not urgently needed. However, in the long run, the four reference populations of the HapMap project, although serving as a good starting point, may not be sufficient to achieve the goal of the project. One might argue that populations such as Druze, Khanty, etc do not represent our species, and thus the analyses in these populations would be too specific to serve a general purpose. However, to establish a final and a clear link among modern human populations, these populations are far from negligible. Ultimately, in both genomic and evolutionary contexts, we will need information from them. Isolated populations are often considered ideal for studies of complex diseases because the genetic component is expected to be less complex.[34] Our data argue that it is precisely those populations that are least likely to be similar to an existing HapMap population. Diverse populations along with a large number of high-density markers are needed to assure broad applicability of the results. So, Francis Collins (*The Scientist*, June 30, 2003) was being somewhat optimistic in his assessment when he said.

> We may be OK without sampling very broadly throughout the world. The similarities are substantial enough that maybe three [geographic] areas will be sufficient to produce a tool that you can use anywhere.

Our goal in this study is to examine the variation in LD among a global sampling of populations and its consequences for selecting tagSNPs. We find significant variation in LD patterns among populations, both among and within geographic regions, suggesting that the data on existing HapMap are insufficient for studies in some other populations. We have not considered other potential inadequacies of the HapMap effort.[35] The variation we have documented is a significant factor in some efforts to study the variation underlying disease phenotypes. However, because of the historic subsetting of variation as humans spread around the world, there is a clear but imperfect asymmetric pattern of tagSNP transferability from 'older' to 'newer' populations. Our studies show that it is advisable to assess a population of interest for genetic similarity to a HapMap population instead of simply grouping it according to geographic location. Our analyses also show that one may be able to compensate for dissimilarity to a HapMap population by increasing the stringency of tagSNP selection in the most similar HapMap population while expecting a lower coverage in the study population. We see, for example, that tagSNPs chosen at $r^2 = 0.90$ can work at $r^2 = 0.80$ much better than tagSNPs chosen at $r^2 = 0.80$. The increase in number of tagSNPs compensates for uncertainty in transferability.

Acknowledgements

We thank F Black, B Bonne-Tamir, L Cavalli-Sforza, K Dumars, J Friedlaender, D Goldman, E Grigorenko, SLB Kajuna, NJ Karoma, KS Kendler, WC Knowler, S Kungulilo, D Lawrence, R-B Lu, A Odunsi, F Okonofua, F Oronsaye, J Parnas, L Peltonen, LO Schulz, D Upson, D Wallace, KM Weiss, S Williams, OV Zhukova for helping assemble the diverse population collection used in this study. Some cell lines were made available by the Coriell Institute for Medical Research and by the National Laboratory for the Genetics of Israeli Populations. Special thanks are given to the many hundreds of individuals who volunteered to give blood samples for studies. This work was supported in part by NIH GM57672.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Kidd KK, Morar B, Castiglione CM *et al*: A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* 1998; **103**: 211–227.
3 Kidd JR, Pakstis AJ, Zhao H *et al*: Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet* 2000; **66**: 1882–1899.
4 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
5 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
6 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.

7 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.

8 International HapMap Consortium: International HapMap project. *Nature* 2003; **426**: 789–796.

9 International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.

10 Clark AG, Weiss KM, Nickerson DA *et al*: Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 1998; **63**: 595–612.

11 Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF: Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet* 2000; **66**: 69–83.

12 Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 2002; **71**: 1227–1234.

13 González-Neira A, Calafell F, Navarro A *et al*: Geographic stratification of linkage disequilibrium: a worldwide population study in a region of chromosome 22. *Hum Genomics* 2004; **1**: 399–409.

14 González-Neira A, Ke X, Lao O *et al*: The portability of tagSNPs across populations: a worldwide survey. *Genome Res* 2006; **16**: 323–330.

15 Sawyer SL, Mukherjee N, Pakstis AJ *et al*: Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* 2005; **13**: 677–686.

16 Liu N, Sawyer SL, Mukherjee N *et al*: Haplotype block structures show significant variation among populations. *Genet Epidemiol* 2004; **27**: 385–400.

17 Tishkoff SA, Kidd KK: Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* 2004; **36** (Suppl 11): S21–S27.

18 Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK: ALFRED: an allele frequency database for diverse populations and DNA polymorphisms – an update. *Nucleic Acids Res* 2001; **29**: 317–319.

19 Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK: ALFRED: an allele frequency database for anthropology. *Am J Phys Anthropol* 2002; **119**: 77–83.

20 Wright S: *Evolution and the Genetics of Populations. Vol 2: The Theory of Gene Frequencies*. University of Chicago Press: Chicago IL, 1969.

21 Gu S, Pakstis AJ, Kidd KK: HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* 2005; **21**: 3938–3939.

22 Wang N, Deng M, Chen T, Waterman MS, Sun F: A dynamic programming algorithm for haplotype partitioning. *Proc Natl Acad Sci USA* 2002; **99**: 7335–7339.

23 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.

24 Clayton D##http://www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf, 2001.

25 Johnson GCL, Esposito L, Barratt BJ *et al*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.

26 Nothnagel M, Furst R, Rhode K: Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered* 2003; **54**: 186–198.

27 de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.

28 de Bakker PIW, Graham RR, Alshuler D, Henderson BE, Haiman CA: Transferability of Tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. *Pacific Symp Biocomput* 2006; **11**: 478–486.

29 Abecasis GR, Cookson WO: GOLD – graphical overview of linkage disequilibrium. *Bioinformatics* 2000; **16**: 182–183.

30 Wall JD, Pritchard JK: Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet* 2003; **73**: 502–515.

31 Kidd KK, Pakstis AJ, Speed WC, Kidd JR: Understanding human DNA sequence variation. *J Hered* 2004; **95**: 406–420.

32 Hey J: On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol* 2005; **3**: e193.

33 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization Algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–959.

34 Shifman S, Darvasi A: The value of isolated populations. *Nat Genet* 2001; **28**: 309–310.

35 Terwilliger JD, Hiekkalinna T: An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur J Hum Genet* 2006; **14**: 426–437.

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