Human Molecular Genetics, 2004, Vol. 13, No. 21 **2557–2565**
doi:10.1093/hmg/ddh294
Advance Access published on September 14, 2004

Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples

Xiayi Ke[1], Caroline Durrant[1], Andrew P. Morris[1], Sarah Hunt[2], David R. Bentley[2], Panos Deloukas[2] and Lon R. Cardon[1,*]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK and [2]Wellcome Trust Sanger Institute, Hinxton, UK

Received June 18, 2004; Revised August 17, 2004; Accepted September 8, 2004

Haplotype tagging is a means of retaining most of the information in high density marker maps, while reducing genotyping requirements. Estimates of the numbers of tagging SNPs required to cover the human genome have varied widely, ranging from 100 000 to 1 000 000. Tagging has been applied to a number of gene-based datasets but has not been evaluated in contexts reflecting those of genome-wide association studies—large chromosome regions and multiple samples drawn from the same population. We analysed 5000 common markers across a 10 Mb segment of human chromosome 20 in three samples (UK Caucasian, CEPH Caucasian, African American) to evaluate tagging efficiency and consistency. Overall, the results indicate a high degree of efficiency, yielding 3–5-fold savings in Caucasians and 2–3-fold savings in African Americans. These levels varied according to linkage disequilibrium (LD) levels, tagging thresholds and allele frequencies, but in high LD regions they did not vary markedly due to marker density. However, a strong positive relationship between marker density and tagging was observed, relating to the fact that increasing marker density yields greater sequence coverage in high LD, thus requiring more tag SNPs to cover a greater fraction of the genome. Encouragingly, whatever the density employed, a high level of robustness was observed between UK and CEPH samples, as most of the htSNPs selected in one sample were also appropriate as tags in the other.

INTRODUCTION

High-density sets of well-characterized genetic markers are emerging for complete human chromosomes or large contiguous regions (1,2) (http://www.hapmap.org/downloads/encode1.html.en), and have been assembled for a number of specific genes (3–6). It is expected that validated SNP maps of 1 marker/5 Kb will soon encompass the entire human genome as part of the international HapMap project (7). By providing a very large set of validated SNPs in multiple populations, the HapMap project aims to benefit indirect association studies across a wide-range of complex traits.

When markers are correlated in the population (i.e. in linkage disequilibrium, LD), redundant information exists in the sense that the observed genotypes at one marker yield information about those at another. In the most extreme case, two SNPs may be perfectly correlated in the population (the r^2 LD measure = 1.0), so that each individual's genotype at one SNP is completely determined by that at the other. In this case, there is nothing to be gained by genotyping both SNPs, as either will suffice. In less extreme cases, thresholds can be chosen so that a lower value of the r^2 coefficient is deemed acceptable or that specific markers are selected to predict a subset of all observed haplotypes. In general, selection of non-redundant markers from a larger set has been called 'haplotype tagging', and the resulting SNPs selected are referred to as haplotype-tag SNPs (htSNPs) or just tag-SNPs (tSNPs) (8,9). Given an initial set of densely spaced and potentially redundant markers, haplotype tagging aims to reduce the scale and cost of genotyping in subsequent applications, yet maintain most or all of the information provided in the dense map.

A number of different algorithms and computer programmes have been developed to define sets of htSNPs (1,8,10–19) and it seems likely that these developments will continue (comparisons and reviews in 9,12). Applications of these approaches have generally focussed on genes (8,12,15,16,19–22), small chromosome regions (9,23,24) and a large

*To whom correspondence should be addressed at: Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. Tel: +44 1865287591; Fax: +44 1865287697; Email: lon@well.ox.ac.uk

set of SNPs genotyped on a small number of individuals (1,10). The results consistently point to efficiencies for tagging common variants. Although there is substantial locus-specific variability, the findings suggest an average savings of 2–5-fold in Western European populations (9,19,20). As a result, development of tagging marker sets is a key objective of the HapMap project (7).

Despite the relatively large number of gene- and region-based analyses of tagging SNPs, tagging characteristics have not yet been evaluated in densely spaced SNPs covering large chromosome regions in multiple populations, although this is the real context of many trait association studies. Questions concerning chromosome-wide tagging efficiency, marker density effects and consistency of results within populations are of direct relevance for practical uses of haplotype tagging. Here, we investigate these issues using approximately 5000 SNPs genotyped at a density of 1 SNP/2.3 Kb along a 10 Mb contiguous segment of chromosome 20q12–13.2 (25). The samples genotyped include 96 unrelated Caucasians from the UK, 48 CEPH founders and 97 unrelated African Americans (2). Apart from providing a dense set of SNPs across a large region, a unique feature of this dataset is the existence of two samples of Western European ancestry (UK Caucasians and CEPH founders). Comparing the consistency and robustness of haplotype tagging in these two samples mirrors the expected use of tag SNPs in disease gene studies, i.e. identifying htSNPs in one sample and applying them in another sample drawn from the same population.

RESULTS

General levels of tagging efficiency

The overall efficiency of haplotype tagging across all high LD regions of chromosome 20q12–q13.2 is shown in Figure 1, using the full set (1 SNP/2.3 Kb) of markers of UK Caucasian and African American samples. The efficiency is shown for situations in which htSNPs were selected to explain all haplotypes in the region, no matter how rare (labelled '100% diversity'), versus those in which htSNPs were required to account for 80% of the haplotype diversity.

The results in Figure 1 indicate several main trends. First, summing over the entire 10 Mb region, tagging all haplotypes in high LD regions yields reductions of ~3-fold in Western European samples and 2-fold in African American samples. These savings are of similar magnitude of those described in gene-based applications of tagging (9,20). Second, the efficiency is notably reduced in samples of African ancestry, owing to the well-known LD differences between African populations and those from Western Europe (26). Finally, relaxing the htSNP requirement so that only 80% of the initial haplotypes are tagged yields a dramatic improvement over the requirement of explaining all haplotype diversity. The tagging efficiencies are nearly twice as high for 80% diversity than for 100% diversity, in both Western European samples and African American samples.

More detailed information about tagging efficiency is presented in Table 1, in which each region to be tagged is considered in terms of the number of markers in the initial genotyping set. With high-density panels, most regions of consistently high LD

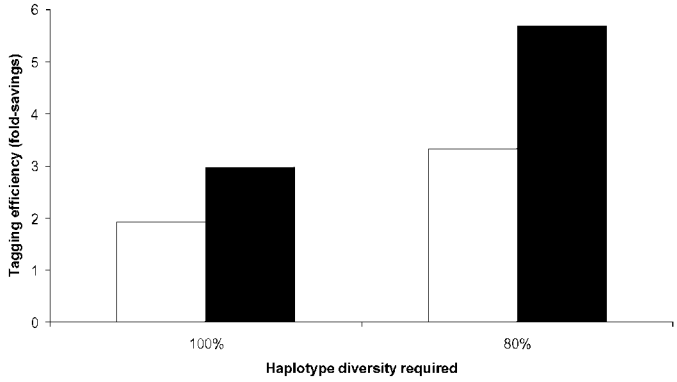

Figure 1. Haplotype tagging efficiency in a dense marker map. Efficiency is defined as the total number of markers genotyped divided by the number of htSNPs required to explain all haplotypes (100% diversity) or the most common haplotypes (80% diversity). The results indicate average efficiencies obtained across all high LD regions in the 1 SNP/2.3 Kb marker set on chromosome 20. Dark bars denote UK Caucasian sample and light bars depict African American sample.

are short and involve relatively few markers (2,6). This trend is reflected in the values in the columns labelled '*N* regions', where it may be seen that for this 2 kb map most of the regions contain less than 10 markers, although some can be exceptionally long (regions of 40 or more markers imply an LD tract of >80 Kb in length). The short regions are naturally less efficient for tagging, as there are relatively few markers from which to select htSNPs. In contrast, the long regions offer substantial savings from tagging, reaching 12-fold reduction for the longest high LD segments observed. In some regions of the genome, as few as four markers may be needed to capture most of the variability in a segment of 80–100 Kb.

Although haplotype reconstruction can be inaccurate and tagging as a cost saving strategy is largely doubtful for low LD regions, similar analyses were carried out for the low LD regions anyway as a comparison and also for random regions (regions without regard to LD) in the 10 Mb segment. As expected, tagging in the low LD regions was less efficient than that in random regions, which in turn was lower than that in high LD regions. However, the trend, i.e. the dependency of tagging efficiency on number of markers in a region, was similar in all the three different kinds of regions.

Effects of allele frequencies

Figure 2 shows the haplotype r^2 values between htSNP haplotypes selected from a 5 Kb map and the unobserved or 'hidden SNPs' of a 2.3 Kb map. Hidden SNPs in the 2.3 Kb map, which had minor allele frequencies >20% were predicted very well by htSNPs drawn from the 5 Kb map (haplotype $r^2 > 0.85$). However, the correlation with SNPs having rarer alleles dropped rapidly, particularly in the case of htSNPs initially chosen to explain 80% haplotype diversity, where the values dropped to 0.75 and 0.59 for alleles of 10–20% and <10% frequency, respectively. These results emphasize the difficulties in tagging SNPs with rare alleles, as discussed by Weale *et al.*

It might be expected that optimal htSNPs would have the same allele frequency profile as the hidden SNPs they are

Table 1. Tagging efficiency as a function of the number of SNPs in the region to tag

Haplotype diversity explained by htSNPs	*N* SNPs in region	UK Caucasian		African American	
		N regions	htSNP efficiency	*N* regions	htSNP efficiency
High LD regions					
100%	2–10	201	1.60	277	1.33
	10–20	71	3.04	56	1.94
	20–30	19	4.01	20	2.68
	30–40	12	6.18	6	3.38
	≥40	15	8.85	13	5.29
80%	2–10	201	3.04	277	2.41
	10–20	71	6.49	56	3.24
	20–30	19	7.95	20	4.32
	30–40	12	11.35	6	5.76
	≥40	15	11.96	13	6.16
Low LD regions					
100%	2–10	15	1.22	35	1.14
	10–20	4	2.45	16	1.40
	20–30	0		1	1.62
80%	2–10	15	1.68	35	1.59
	10–20	4	3.18	16	1.61
	20–30	0		1	1.75
Random regions					
100%	2–10	201	1.53	277	1.26
	10–20	71	2.50	56	1.74
	20–30	19	3.90	20	2.41
	30–40	12	4.85	6	3.16
	≥40	15	6.11	13	4.62
80%	2–10	201	2.87	277	2.18
	10–20	71	4.37	56	2.41
	20–30	19	6.67	20	2.70
	30–40	12	7.52	6	4.00
	≥40	15	9.95	13	5.12

Random regions were created whose number and size distribution were the same as the high LD regions.

meant to tag, as is well established for marker allele versus disease allele frequencies in association studies (27–31). Average allele frequency differences between htSNPs and their hidden SNPs are inlaid in Figure 2, showing that the bigger the difference between hidden SNPs and the htSNPs selected to tag them, the poorer the predictive power. In general, however, this effect is unidirectional. Situations in which hidden SNPs have higher average MAF than the htSNPs in the region show relatively slight decreases in haplotype r^2, but those in which the hidden SNPs are rarer than the tagging SNPs show steep haplotype r^2 reductions. This MAF deviation effect is also related to the haplotype diversity required by htSNP sets, with the relaxed set of htSNPs explaining 80% haplotype diversity having larger allele frequency effects than the htSNPs selected to explain 100% diversity. The tagging methodology used here appears to efficiently and robustly capture the information of hidden SNPs which are at least as common as the specific htSNPs selected, but it performs poorly for those which are less frequent in the population. As might be expected of many tagging approaches, the allele frequency of the SNPs selected for tagging is highly correlated with the frequencies of those in the region to be tagged (Fig. 2A).

There are two scenarios observed in the data which are relevant to the relationship between rare SNPs and htSNP sets. A rare SNP could be associated with a specific haplotype defined by htSNPs, in which case the haplotype r^2 would be 1.0. Alternatively, it could subdivide a particular haplotype and, in that case, the haplotype r^2 would be less than 1.0. How much of a drop in haplotype r^2 depended on the allele frequency of the SNP itself as well as the frequency of the haplotype being subdivided, as illustrated by the haplotype r^2 formula (see Materials and Methods). When htSNPs were selected requiring 100% haplotype diversity, only 27% of rare hidden SNPs (minor allele frequency $<10\%$) fell into the second scenario. When 80% haplotype diversity was required instead, 100% of such SNPs were found to be subdividing rather than associating with a particular haplotype. Therefore, the savings in genotyping achieved by reducing the required haplotype diversity had a significant impact on detecting rare causal SNPs because rare haplotypes were usually ignored.

It has been suggested that tagging SNPs selected on the basis of haplotype diversity or bins do not perform better than randomly selected markers in retaining power (31). To investigate whether this was true with our sample data in the 10 Mb segment, we created five random marker sets at the 1 SNP/5 Kb map (i.e. for each high LD region, the same

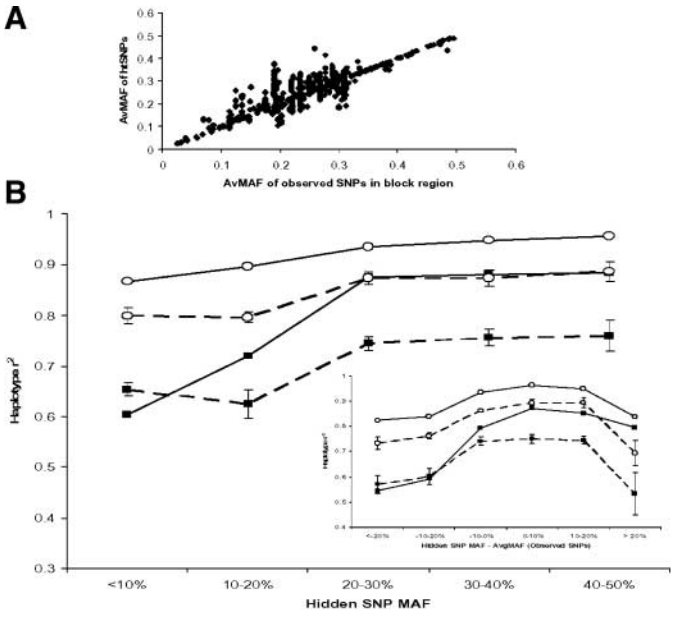

Figure 2. Relationship of allele frequency patterns to tagging efficiency. (**A**) Scatterplot between average minor allele frequencies (AvMAF) of known markers in high LD regions and AvMAF of their individual htSNP sets. The results show the allele frequency correspondence for htSNPs selected at a 1 SNP/5 Kb density and required to explain 100% of the haplotype diversity. (**B**) Average haplotype r^2 for all hidden SNPs as a function of their minor allele frequency. The inset lines show the relationship between haplotype r^2 and the average difference in allele frequency between the hidden SNP and the observed SNPs. For high LD regions, htSNPs selected at the 1 SNP/5 Kb marker set, with required haplotype diversity at 100% (solid lines with cycled data points) or 80% (solid lines with squared data points), were applied to the marker set of 1 SNP/2.3 Kb density. As a comparison, five randomly selected marker sets with equivalent number of markers as the htSNP sets at both 100% (dashed line with circles) and 80% haplotype diversity (dashed line with squares) were also created for each high LD region at the 1 SNP/5 Kb density. These random markers were also applied to the marker set of 1 SNP/2.3 Kb density in the same way as htSNPs.

number of random markers as the tagging SNPs was selected). We chose haplotype r^2 as the evaluation measure to utilize haplotype information in a similar way as in selecting htSNPs. As shown in Figure 2, when 100% diversity was required, htSNPs always performed better than random SNPs. Dropping diversity to 80%, random markers would perform slightly better than htSNPs only if a hidden SNP was rare (minor allele frequency $<10\%$). This was likely because rare haplotypes are excluded when htSNPs were selected, whereas random marker sets could include them by chance. For more common SNPs, higher correlation coefficients were observed with htSNPs than with random markers. This result suggests that tagging based on haplotype diversity is in general a worthwhile effort.

Efficiency and consistency at different marker densities

To explore the effects of marker density on tagging efficiency and consistency, we randomly selected subsets of markers from the full dataset to yield marker panels with densities of 1 marker per 3, 4 and 5 Kb. The results in Figure 3 show

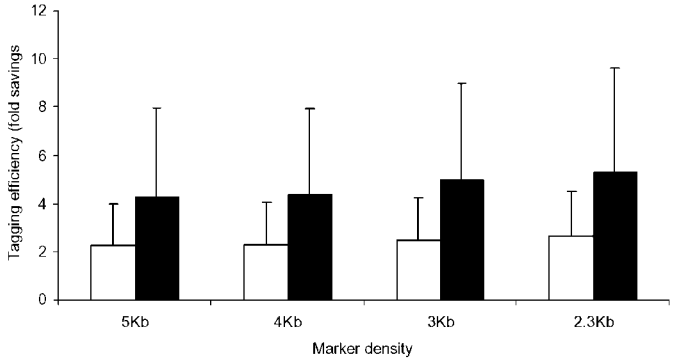

Figure 3. Effects of marker density on tagging efficiency. The main figure shows the tagging efficiency obtained in UK Caucasian samples for subsets of the full dataset which yielded the map densities as shown. Dark bars indicate 100% diversity required and light bars 80% diversity required when tagging SNPs were selected.

the tagging efficiency for the high LD regions of chromosome 20. When all haplotypes are explained by htSNPs, the effects of different marker densities are relatively small; i.e. with sparse maps, there are fewer initial markers genotyped than with dense maps, but the proportion of htSNPs required is approximately the same. In the case of 80% diversity, however, there is a clear gain in efficiency as the density increases. This gain is the result of adding markers in regions of high LD, which only account for a few, if any, rare haplotypes and, therefore, do not contribute enough to the haplotype diversity to be tagged.

Using the same increasing density strategy, we also considered each of the new markers as a 'hidden SNP' and tested the explanatory power of the htSNP sets identified in the initial marker set using haplotype r^2 (12). For every new SNP in the full marker set (1 SNP/2.3 Kb), the haplotype r^2 of the htSNPs identified at an initial density of 1 SNP/5 Kb for the high LD regions was evaluated. For the unrelated Western European samples, when the initial htSNPs were selected to explain 100% of the haplotype diversity in the 5 Kb panel, those same htSNPs revealed an average haplotype r^2 of 0.93 with haplotypes defined by the 2.3 Kb set of markers. When the initial htSNPs were selected to explain 80% of the haplotype diversity, their average haplotype r^2 in the dense set was 0.80. The results for African American data were similar, with respective haplotype r^2 values of 0.91 and 0.79. These results suggest that htSNPs drawn from high LD regions at a 5 Kb density explain nearly as much of the variability in those same regions as those drawn from a map having twice as many markers.

Despite the apparent similarities in haplotype r^2 across densities and the gains in efficiency with increasing density, the number of SNPs required to tag a region increases with marker density (Table 2). In Western Europeans, the number of htSNPs nearly doubles in the 5–2.3 Kb density range (340–619); whereas in African Americans, the effect is even more pronounced, increasing >3-fold (319–970).

To further explore the relationship between LD, marker density and tagging requirements, we applied the tagging approach of Carlson *et al.* (19). This approach relaxes all

Table 2. Number of tagging SNPs required for chromosome 20, with genome-wide extrapolations

Map density (Kb)	Chromosome 20q12–q13.2				Genome-wide extrapolation	
	N markers in region	Sequence in high LD (%)	*N* htSNPs to cover high LD	*N* tag SNPs to cover entire region	*N* htSNPs to cover high LD	*N* htSNPs to cover entire genome
UK Caucasians						
5	2020	52	340	1081	102 000	324 300
4	2525	58	425	1373	127 500	411 900
3	3366	64	498	1911	149 400	573 300
2.3	4337	68	609	2787	182 700	836 100
African Americans						
5	2020	38	319	1398	95 700	419 400
4	2525	42	567	1713	170 100	513 900
3	3366	50	758	2328	227 400	698 400
2.3	4337	56	970	3472	291 000	1 041 600

block requirements and uses only the correlations between all markers, whatever their respective locations. Application of this approach also yielded a substantial increase in the number of markers required according to marker density (Table 2). The number of tag SNPs required by this method is, in part, higher than those from the block-based approach because the former attempts to cover the entire region, whereas the latter focuses only on high LD regions.

As chromosome 20 represents ~2% of the genome (25) and the 10 Mb region represents about 1/6 of chromosome 20, general extrapolations of genome-wide requirements may be obtained by multiplying the chromosome 20 figures by 300. These extrapolations, shown in Table 2, reveal very different estimates of tagging requirements from the high LD versus entire region assessments. The extrapolations suggest that focussing on only high LD would require only about 100 000–300 000 for Western European or African American samples, whatever the density (up to 2.3 Kb), whereas attempting to encompass both high and low LD regions would necessitate about three to four times more markers, nearing 1 million for Western Europeans and exceeding that figure for African Americans. These different figures suggest that the actual number of tag SNPs required depends explicitly on the aims and assumptions of the investigator, as well as the marker density, local LD patterns and tagging methods used.

Using tag SNPs in new samples

The main objective of haplotype tagging is to genotype densely in one sample in order to reduce the genotyping required in subsequent applications. The primary assumption underlying this objective is that the frequency patterns of common haplotypes are similar within populations, and thus htSNPs selected in one sample should provide the same coverage in another sample drawn from the same population. We evaluated the level of consistency in Western Europeans by comparing the properties of htSNPs in our UK Caucasian sample with CEPH founders, all of which are of Western European ancestry. We selected htSNPs from UK Caucasians and then evaluated the degree to which they explain the haplotypes in CEPH founders. Table 3 shows that the haplotype r^2 remains very high in the dense marker panels. For example, htSNPs selected to explain 100% haplotype diversity in UK Caucasian samples explain 96% of the variance in CEPH haplotypes, and htSNPs covering 80% of UK Caucasian diversity explain 73% of that in CEPH data. Similarly, there is a large degree of overlap between the htSNPs chosen in the two samples; ~90% of the htSNPs selected in one sample would also be selected as htSNPs in the other sample. This proportional overlap does not differ between the 100% and 80% diversity requirement. Compared with the high consistency of tagging within West European population, the corresponding figures dropped when the same htSNPs were applied to African American samples, particularly if the diversity was 80% in which the haplotype r^2 dropped to 0.58.

To further explore the generalizability of tagging SNPs, we compared UK and CEPH data for map densities of 2.6, 3, 4, 5 Kb. By selecting htSNPs at a coarse density in UK sample, and then looking at how well they explain the haplotypes derived from a finer SNP map in CEPH data, we can begin to appreciate how much practical information might be gained by extending current maps of ~1 SNP/5 Kb to finer-scale marker densities. Table 4 shows that when LD is high, there is apparently little difference in sample consistency between tagging at a 5 Kb density and tagging at a 2.3 Kb density. The diagonal elements in Table 4 indicate that for htSNPs selected to explain 100% of the diversity in one sample (lower matrix), the haplotype r^2 is at least 95% for another sample genotyped at the same marker density. This level of consistency is similar to that observed within the same sample but at different marker densities (Figs 2 and 3). The off-diagonal elements show haplotype r^2 values in excess of 93% when a coarser map is used to define the htSNPs in another sample. When the htSNPs are selected to explain 80% of the diversity (upper matrix), the loss of explained variance is similarly small, with all haplotype r^2 values at least 0.73.

DISCUSSION

In this study, properties of tagging SNPs were analysed using approximately 5000 SNPs genotyped in multiple samples across a 10 Mb segment of human chromosome 20. Haplotype correlations between tagging SNP sets and ungenotyped

Table 3. Robustness of htSNPs selected in sample of UK unrelateds when applied to CEPH sample

Haplotype diversity explained by htSNPs (%)	*N* SNPs in region	*N* htSNPs in			*N* htSNPs in common		Haplotype r^2	
		CAUC	CEPH	AFAM	CAUC–CEPH	CAUC–AFAM	CAUC–CEPH	CAUC–AFAM
100	2139	843	958	1156	757 (90%)	753 (89%)	0.96	0.84
80	2139	431	438	727	385 (89%)	355 (82%)	0.73	0.58

CAUC denotes UK Caucasians; CEPH denotes CEPH founders and AFAM African Americans.

Table 4. Haplotype r^2 in the CEPH samples for htSNPs selected in UK unrelateds

		CEPH founders				UK unrelated (80% diversity)					
		2 Kb	3 Kb	4 Kb	5 Kb	2 Kb	3 Kb	4 Kb	5 Kb		
UK unrelated	2 Kb	**0.96**				**0.75**	0.75	0.74	0.73	2 Kb	CEPH founders
(100% diversity)	3 Kb	0.95	**0.96**				**0.74**	0.74	0.73	3 Kb	
	4 Kb	0.94	0.94	**0.96**				**0.74**	0.73	4 Kb	
	5 Kb	0.93	0.93	0.94	**0.95**				**0.73**	5 Kb	

SNPs were used as a measure of the reduction in power to detect effects at unassayed SNPs. Several of the present results from our analyses of this large contiguous region are in strong agreement with those previously obtained using smaller marker sets in genes and gene-regions (9,19,20). Others, particularly those relating to marker density and the comparisons of UK and CEPH samples, extend the previous results into the practical context of local gene discovery and eventually genome-wide association.

On average, haplotype tagging offers consistent genotyping savings in regions of high LD, although the level of efficiency varies substantially according to allele frequency of htSNPs versus ungenotyped markers, size and marker coverage of the genomic region and extent of haplotype diversity explained. The efficiency is particularly strong in large genomic tracts of high LD, where the savings occur due to the redundancy of many markers spread over a large distance. The highest level of savings observed on chromosome 20 was 12-fold in UK Caucasians and 6-fold in African Americans. In contrast, small genomic regions or those in which LD was low revealed almost no gains from haplotype tagging.

In most haplotype tagging methods, including those employed here, arbitrary thresholds must be set concerning the level of haplotype diversity, marker correlations or proportion of variance explained by htSNPs. Our results suggest that the exact level of these thresholds may have important implications for the efficiency and robustness of the tagging process. A strict criterion in which htSNPs were required to explain 100% of the haplotype diversity yielded about one-half of the level of savings as a criterion of explaining only 80% of the diversity (3.0- versus 5.7-fold in Caucasians; 1.9- versus 3.3-fold reductions in African Americans). It is not surprising that stricter thresholds demand more genotyping, but the high level of change is relevant for association study design.

Allele frequencies are a key feature of haplotype tagging (12). The present results suggest that the most important feature is not simply whether htSNPs are rare or common, but the extent to which they match the frequencies of ungenotyped (or 'hidden') SNPs. Tagging was much less reliable for untagged SNPs which were rarer than the htSNPs selected than for those that were at least as common as the htSNPs. In addition, the high correlation between htSNPs and the underlying hidden SNPs may have implications for disease gene mapping. For example, if a causal SNP occurs at a frequency that is different from the average of the observed set of markers in a region, the chance of this causal SNP being detected by any of the htSNP sets in the region would be low, especially if the SNP was rare. The closer the frequencies values, the greater the likelihood of detecting the association effects by haplotype tagging.

The primary objective of this study was to evaluate efficiency and reliability of tagging in practical settings. In this regard, the results are encouraging, as different marker densities did not have a serious impact on the efficiency or reliability of haplotype tagging on chromosome 20. We evaluated the effects of marker density on haplotype tagging by choosing subsets of our full (1 SNP/2.3 Kb) marker set to create panels in which SNPs were 3, 4 and 5 Kb apart. The 1 SNP/5 Kb density represents what is proposed for the initial version of the genome-wide haplotype map (7). Nearly 80% of the htSNPs selected at the 1 SNP/5 Kb density map remained htSNPs at the full marker density, a figure which increased when more relaxed levels of explanatory power of htSNPs were allowed. Also, in high LD regions, markers selected at the 1 SNP/5 Kb density explained nearly the same proportion of haplotype variance in a 1 SNP/2.3 Kb map as they were required to explain to meet the initial selection. This holds promise for using tagging strategies to reduce the amount of genotyping effort for common alleles.

A very high degree of consistency was observed between htSNPs selected in our sample of UK Caucasians and US-based CEPH samples. At least 80% of the htSNPs selected in UK Caucasian set would also have been selected as htSNPs in the CEPHs in all situations examined. Moreover, in high LD regions, the proportion of variation explained

by htSNPs in UK Caucasian sample was nearly unchanged when those same SNPs were applied to CEPH data (all differences in haplotype $r^2 \leq 0.07$).

A main outcome from the present study concerns the relationship between marker density and tag SNP selection. The results suggest that the tagging SNPs selected in high LD regions are largely robust to variability in marker density, but at the same time the number of tag SNPs required increases as marker spacing becomes more dense. These findings may be seen as contradictory, since the efficiency suggests greater economy of genotyping whatever the density, whereas the increase in numbers of htSNPs suggests less efficiency. Why is this so? At least one reason for the apparent discrepancy relates to LD. By definition, tagging requires markers to be correlated: higher LD leads to greater efficiency. However, with sparse maps, less of the genome is in high LD than with denser maps. On chromosome 20, only 52% of the 10 Mb segment is in high LD when evaluated using a 5 Kb map, but the high LD sequence coverage is ~70% with a 2.3 Kb map (Table 2). Thus, fewer markers are required with sparse maps, but less of the chromosome is tagged, and consequently, as the density increases, more markers are required to encompass a greater fraction of the chromosome.

Extrapolations from the chromosome 20 data to the number of htSNPs required to cover the human genome highlight the density/LD relationship. If one is interested only in high LD regions, only 100 000–300 000 markers may be required. However, if the tag SNPs are selected from a 5 Kb map, only ~50% of the genome will be covered, since that is roughly the fraction of the genome in high LD at 5 Kb density. Using higher density markers clearly improves the coverage, but at even a 1 SNP every 2 Kb only ~70% of chromosome 20 is covered in our sample of Western Europeans, and less so in African Americans. For complete or near-complete coverage of the genome, including both high and low LD regions, three to four times more SNPs may be needed than that for the high LD regions only, i.e. ~1 000 000 or more selected genetic markers. Interestingly, the lower estimates are consistent with some previous suggestions (9), whereas the higher estimates are in line with other reports (19,20). It thus seems possible that one reason for the apparent discrepancy in tagging estimates is simply different aims of genomic coverage and assumptions about LD patterns by different investigators. Our chromosome 20 data are consistent with both estimates when taking into account the different underlying study objectives.

Although the present results are generally encouraging, it is important to emphasize that certain important questions remain that were not possible to address in the present dataset. First, consideration of rare alleles is clearly a key issue for haplotype tagging. Our results point to some important consequences of allele frequency, but we cannot formally assess the effects below the range of ~4% minor allele frequency due to the genotyping and SNP ascertainment strategies used. Second, our sample sizes are generally small, though typical of those often examined in LD and tagging assessments. It will be important to evaluate the robustness of the LD patterns and the htSNP selections in samples of the size which is likely to be used in practical disease applications, i.e. thousands of individuals. Third, as one of the shorter chromosomes in the human genome, chromosome 20 has a higher overall recombination rate than the genome average (32). Accordingly, the levels of LD are lower and the total sequence covered in high LD is less than some other parts of the genome. Thus, the genome-wide extrapolations from these data may reflect upper bounds on the estimates. Finally, UK and US (CEPH) samples examined here are likely to be more homogeneous than those in the general population from which they were drawn. Evaluating tagging consistency in more representative samples would greatly assist in evaluating the practical utility of tagging for large scale association studies.

MATERIALS AND METHODS

Samples and SNP genotyping

DNA was used from a panel of 96 UK Caucasian and 97 African American individuals. Details of the samples and SNP genotyping are provided in the literature (2). A total of 4427 and 4938 markers were genotyped across a 10 Mb region of chromosome 20 (contig NT_011362.7:3 726 000–13 824 000 bp) in UK and African American samples, respectively. In total, 4337 of the SNPs were common to both populations, yielding an average density of 1 marker/2.3 Kb. In order to evaluate the within-sample tagging consistency, we also examined 48 CEPH founders genotyped on 5324 markers, of which 3810 were in common with the above two population samples (1 marker/2.6 Kb).

Many of the variant sites assessed here were initially identified via resequencing of four individuals, yielding a slightly greater proportion of rare marker alleles than typically found in public databases. However, the genotyping technology employed and the final marker selection protocol resulted in a (largely) uniform distribution of allele frequencies, which under-represents rare alleles in the human genome (33,34). Allele frequency and marker spacing distributions, as well as the LD patterns amongst the SNPs, are described by Ke *et al.* (2).

Characterization of LD

The levels and patterns of LD are a critical feature of any htSNP selection scheme and there are a variety of ways to delimit regions for tagging (e.g. on the basis of genes, LD, number of markers, physical spacing, recombination rates, etc.). For simplicity in the present analysis, we dichotomize the data into high LD segments versus everything else. We do this in order to examine tagging properties in one of the most optimistic scenarios (high LD) to gain a sense of the best possible results, and then contrast this with situations in which the marker correlations are more variable. We defined high LD regions according to the haplotype block definition of Gabriel *et al.* (35), though we note that tagging does not require target regions to be delimited by blocks or even by exceptionally high LD (8,9,19). This approach is expected to yield high tagging efficiencies within blocks since by definition they have high correlations amongst component markers. We would expect that different definitions of high LD could give different absolute results,

but we might anticipate similar trends and general extrapolations. In contrast, low LD regions outside blocks are less amenable to tagging and less accurate for haplotype estimation in population data (36).

Tagging SNP selection

For each high LD region partitioned by the block assessments, haplotypes were estimated using snphap (http://www-gene.cimr.cam.uk/clayton/software). For general htSNP selection, haplotypes with frequency <0.01 were excluded and then tagging SNPs were identified using the SNPtagger programme (13).

For all block-based analyses, two sets of htSNPs were constructed and evaluated. In the first case, htSNPs were selected to explain 100% of the haplotype diversity given the markers analysed. This strict definition defines a baseline which is inefficient but as robust as possible. It also provides a setting in which the specific tagging methodology employed is expected to be largely irrelevant, since different genotype-based tagging approaches would likely have similar requirements to explain all variability in haplotypes (19). As most tagging approaches aim to capture $<100\%$ of the underlying information, we also employed a strategy of tagging of 80% of the haplotype diversity.

In order to obtain a block-independent view of the overall number of SNPs required to tag a region, we also used the greedy algorithm of Carlson *et al.* (19), in which markers exceeding an arbitrary LD level are collected in bins and tag SNPs then selected from within each bin. This method provides an indication of the number of tag SNPs required for any genomic region, which may encompass both high and low LD runs. Following the suggestions of Carlson *et al.*, we set the LD threshold at $r^2 = 0.5$ and used their LDselect.pl script to identify common tSNPs (minor allele frequency ≥ 0.05).

Measures of tagging consistency and efficiency

Following Weale *et al.* (12) and Goldstein *et al.* (9), we use the haplotype r^2, which reflects the degree to which a tagging SNP set explains variability in the haplotypes it is chosen to tag. For the ith tagged SNP in question, this measure is calculated as

$$r_i^2 = 1 - \frac{2m^2 \sum_g f_{gi}(h_g - f_{gi})/h_g}{2m^2 f_i(1 - f_i)},$$

where m is the total number of chromosomes observed, f_i is the frequency of allele '1' at locus i, f_{gi} is the frequency of that allele on the gth haplotype and h_g is the haplotype frequency of the gth htSNP-defined group.

To assess the savings in genotyping offered by tagging, we define 'tagging efficiency' as n/n_h, where n_h is the number of htSNPs selected to cover the region. Note that n is the total number of markers genotyped in this study; it does not reflect any unascertained SNPs as we do not attempt to draw inferences about marker densities finer than those presently available. Regions that fell between two high LD regions and contained only one marker were excluded from the study, which counted for $<5\%$ of the total markers in the whole segment.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust and by grant NEI-12562 from the NIH.

REFERENCES

1. Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
2. Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris, A.P., Bentley, D. *et al.* (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.*, **13**, 577–588.
3. Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E. and Sing, C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.*, **19**, 233–240.
4. Nickerson, D.A., Taylor, S.L., Fullerton, S.M., Weiss, K.M., Clark, A.G., Stengard, J.H., Salomaa, V., Boerwinkle, E. and Sing, C.F. (2000) Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res.*, **10**, 1532–1545.
5. Subrahmanyan, L., Eberle, M.A., Clark, A.G., Kruglyak, L. and Nickerson, D.A. (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am. J. Hum. Genet.*, **69**, 381–395.
6. Crawford, D.C., Carlson, C.S., Rieder, M.J., Carrington, D.P., Yi, Q., Smith, J.D., Eberle, M.A., Kruglyak, L. and Nickerson, D.A. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.*, **74**, 610–622.
7. Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ang, L., Huang, W., Liu, B., Shen, Y. *et al.* (2003) The International HapMap project. *Nature*, **426**, 789–796.
8. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
9. Goldstein, D.B., Ahmadi, K.R., Weale, M.E. and Wood, N.W. (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.*, **19**, 615–622.
10. Zhang, K., Sun, F., Waterman, M.S. and Chen, T. (2003) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am. J. Hum. Genet.*, **73**, 63–73.
11. Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.
12. Weale, M.E., Depondt, C., Macdonald, S.J., Smith, A., Lai, P.S., Shorvon, S.D., Wood, N.W. and Goldstein, D.B. (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.*, **73**, 551–565.
13. Ke, X. and Cardon, L.R. (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, **19**, 287–288.
14. Meng, Z., Zaykin, D.V., Xu, C.F., Wagner, M. and Ehm, M.G. (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.*, **73**, 115–130.

15. Sebastiani, P., Lazarus, R., Weiss, S.T., Kunkel, L.M., Kohane, I.S. and Ramoni, M.F. (2003) Minimal haplotype tagging. *Proc. Natl Acad. Sci. USA*, **100**, 9900–9905.
16. Cousin, E., Genin, E., Mace, S., Ricard, S., Chansac, C., del Zompo, M. and Deleuze, J.F. (2003) Association studies in candidate genes: strategies to select SNPs to be tested. *Hum. Hered.*, **56**, 151–159.
17. Wiuf, C., Laidlaw, Z. and Stumpf, M.P. (2003) Some notes on the combinatorial properties of haplotype tagging. *Math. Biosci.*, **185**, 205–216.
18. Barratt, B.J., Payne, F., Rance, H.E., Nutland, S., Todd, J.A. and Clayton, D.G. (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.*, **66**, 393–405.
19. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
20. Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.*, **56**, 18–31.
21. Goldstein, D.B., Tate, S.K. and Sisodiya, S.M. (2003) Pharmacogenetics goes genomic. *Nat. Rev. Genet.*, **4**, 937–947.
22. Lowe, C.E., Cooper, J.D., Chapman, J.M., Barratt, B.J., Twells, R.C., Green, E.A., Savage, D.A., Guja, C., Ionescu-Tirgoviste, C., Tuomilehto-Wolf, E. *et al.* (2004) Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun.,* **5**, 301–305.
23. Zhang, K., Qin, Z.S., Liu, J.S., Chen, T., Waterman, M.S. and Sun, F. (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.*, **14**, 908–916.
24. Schulze, T.G., Zhang, K., Chen, Y.S., Akula, N., Sun, F. and McMahon, F.J. (2004) Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum. Mol. Genet*., **13**, 335–342.
25. Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Bagguley, C.L. *et al*. (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414**, 865–871.
26. Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1994) *History and Geography of Human Genes*. Princeton University Press, Princeton.
27. Zondervan, K.T. and Cardon, L.R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.*, **5**, 89–100.
28. Abecasis, G.R., Cookson, W.O. and Cardon, L.R. (2001) The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am. J. Hum. Genet.*, **68**, 1463–1474.
29. Muller-Myhsok, B. and Abel, L. (1997) Genetic analysis of complex diseases. *Science*, **275**, 1328–1329.
30. Tu, I.P. and Whittemore. A.S. (1999) Power of association and linkage tests when the disease alleles are unobserved. *Am. J. Hum. Genet.*, **64**, 641–649.
31. Zhang, W., Collin, A. and Morton, N. (2004) Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum. Genet.,* **115**, 157–164.
32. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al*. (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 242–247.
33. Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, **3**, 299–309.
34. Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L. and Nickerson, D.A. (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.*, **33**, 518–521.
35. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel. B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
36. Fallin, D. and Schork, N.J. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.*, **67**, 947–959.