European Journal of Human Genetics (2007) 15, 1063–1070

www.nature.com/ejhg

ARTICLE

Tag SNP selection for candidate gene association studies using HapMap and gene resequencing data

Zongli Xu[1], Norman L Kaplan[2] and Jack A Taylor*[,1,3]

[1]*Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA;*
[2]*Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA;*
[3]*Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA*

HapMap provides linkage disequilibrium (LD) information on a sample of 3.7 million SNPs that can be used for tag SNP selection in whole-genome association studies. HapMap can also be used for tag SNP selection in candidate genes, although its performance has yet to be evaluated against gene resequencing data, where there is near-complete SNP ascertainment. The Environmental Genome Project (EGP) is the largest gene resequencing effort to date with over 500 resequenced genes. We used HapMap data to select tag SNPs and calculated the proportions of common SNPs (MAF $\geq$ 0.05) tagged ($\rho^2 \geq 0.8$) for each of 127 EGP Panel 2 genes where individual ethnic information was available. Median gene-tagging proportions are 50, 80 and 74% for African, Asian, and European groups, respectively. These low gene-tagging proportions may be problematic for some candidate gene studies. In addition, although HapMap targeted nonsynonymous SNPs (nsSNPs), we estimate only $\sim$30% of nonsynonymous SNPs in EGP are in high LD with any HapMap SNP. We show that gene-tagging proportions can be improved by adding a relatively small number of tag SNPs that were selected based on resequencing data. We also demonstrate that ethnic-mixed data can be used to improve HapMap gene-tagging proportions, but are not as efficient as ethnic-specific data. Finally, we generalized the greedy algorithm proposed by Carlson *et al* (2004) to select tag SNPs for multiple populations and implemented the algorithm into a freely available software package *mPopTag*.
European Journal of Human Genetics (2007) **15,** 1063–1070; doi:10.1038/sj.ejhg.5201875; published online 13 June 2007

Keywords: Environmental Genome Project; HapMap; gene resequencing; tag SNP; ethnic-mixed data; composite LD

Introduction

The International HapMap Project has detailed information on genetic variation across the genome.[1] An important use of these data is to help identify genetic determinants of disease. HapMap Release 20 has genotype data for more than 3.7 million SNPs for several populations (http://www.hapmap.org/). Simulations with HapMap ENCODE (Encyclopedia of DNA Elements) Project data, (resequencing of 10 500-kb genomic regions in 48 individuals and subsequent genotyping of all discovered SNPs as well as all SNPs in dbSNP at the time in the 270 HapMap DNA samples), estimated that 94% of the common SNPs (minor allele frequency, MAF $\geq$ 0.05) in non-African populations and 81% in Yoruba from Ibadan, Nigeria (YRI) populations are in high linkage disequilibrium (LD) with at least one of the SNPs in HapMap.[1] These simulations suggest that HapMap SNP density may be adequate for whole-genome association studies.

***Correspondence: Dr JA Taylor, Epidemiology Branch, National Institute of Environmental Health Sciences, MD A3-05, 111 Alexander Drive, PO box 12233, Research Triangle Park, NC 27709, USA.**
Tel: +1 919 541 4631; Fax: +1 919 541 2511;
E-mail: taylor@niehs.nih.gov
Received 12 July 2006; revised 30 March 2007; accepted 11 May 2007; published online 13 June 2007

Investigators are also using HapMap data for SNP selection in candidate gene association studies.[1,2] Because HapMap collects samples from SNPs that have been deposited into dbSNP (http://www.ncbi.nlm.nih.gov/SNP), it only has partial information on gene polymorphisms, whereas near-complete ascertainment of common SNPs in genes can be obtained through gene resequencing. The largest gene resequencing effort to date is the Environmental Genome Project (EGP) sponsored by National Institute of Environmental Health Science[3] (http://www.niehs.nih.gov/envgenom/home.htm), which at the time of this study has resequenced 518 genes in 90 to 95 people of different ethnic backgrounds and has identified more than 70 000 SNPs. In total, EGP has resequenced more than 12 Mb of the human genome, although individual ethnic information is available only for 127 genes resequenced in EGP Panel 2. We used HapMap data to identify tag SNPs for each of these 127 genes and then, using the catalog of common SNPs identified through EGP resequencing, we estimated gene-tagging proportions of HapMap tag SNPs in each of three ethnic groups. In addition, we considered strategies to improve gene-tagging proportions beyond those obtained using HapMap tag SNPs.

The 391 genes resequenced in EGP Panel 1 used 90 individuals drawn from the ethnically diverse Polymorphism Discovery Resource.[4] Because of ethical, legal, and social implications (ELSI), ethnic identifiers were removed, resulting in an ethnic-mixed sample, but one with known ethnic proportions. The utility of tag SNPs chosen from an ethnic-mixed sample is unclear, because allele frequencies and/or underlying LD patterns may differ between populations,[5] We investigated this problem by using ethnic-pooled data for EGP Panel 2 genes for which we have individual ethnic data.

Candidate gene studies often include individuals from multiple ethnic groups, which may require the use of different ethnic-specific panels of tag SNPs. It would be reasonable and certainly more convenient to have one set of tag SNPs that can be used in multiple populations. Similar in purpose to the TagIT[6] and MultiPop-TagSelect[7] methods, we generalized the greedy algorithm proposed by Carlson *et al* (2004)[21] to select tag SNPs for multiple populations. We used this algorithm to choose HapMap multipopulation tag SNPs and evaluated gene-tagging proportions for EGP Panel 2 genes.

Because nonsynonymous coding SNPs (nsSNPs) are a high priority for candidate gene-association studies,[8,9] HapMap made a special effort to include as many nsSNPs as possible.[1] Despite HapMap's effort, its information on nsSNPs may be limited, because most nsSNPs have low MAF.[8–11] To quantify HapMap's success in capturing nsSNPs, we used EGP resequence data to estimate the fraction of nsSNPs that are either in HapMap or in high LD with a SNP in HapMap.

Materials and methods

Data

The EGP selected for resequencing those genes thought to be involved in susceptibility to environmentally associated disease. The major focus of this effort was on genes associated with DNA repair, cell cycle regulation, apoptosis, and metabolism. These genes are widely distributed across all chromosomes, except for the Y chromosome. At the time of this study, genotypes based on resequencing data were available from the EGP website for 52 387 SNPs in 391 genes from EGP Panel 1 and for 18 850 SNPs in 127 genes from EGP Panel 2. By examining the date of deposit, we found 52 352 (73%) of the SNPs in Panel 1 and 2 were novel at the time of their deposit into dbSNP.[12] Approximately 17% of the novel SNPs were common (MAF $\geq$ 0.05) in EGP data. We considered only biallelic SNPs with less than 20% missing genotype data, resulting in 48 697 SNPs in EGP Panel 1 and 17 495 SNPs in EGP Panel 2. The EGP resequencing effort applied a number of measures to assure data quality and had an average base call Phred score >45 (99.998% accuracy of the base call).[13]

EGP Panel 1 has DNA from 90 individuals, that includes 24 African-Americans, 24 Asian-Americans, 24 European-Americans, 12 Hispanic-Americans, and 6 Native-Americans, with equal numbers of males and females drawn from the Polymorphism Discovery Resource.[4] EGP Panel 2 has DNA from an independent set of 95 individuals (http://egp.gs.washington.edu/), that includes 15 African-Americans (AA), 12 YRI, 12 Japanese in Tokyo, Japan (JPT), 12 Han Chinese in Beijing, China (CHB), 22 CEPH (Utah residents with ancestry from northern and western Europe) (CEPH) and 22 Hispanics (HISP). Fifty-eight of the individuals (12 YRI, 12 JPT, 12 CHB, and 22 CEPH) in EGP Panel 2 were also included in HapMap. Although African-Americans have an admixed ancestry,[4] a recent study has shown that the LD pattern of African-Americans was similar to YRI,[14] and therefore, we combined the two groups as 'African'. Similarly, Chinese and Japanese data were combined as 'Asian'. To mimic the EGP Panel 1 ethnic-mixed sample, we also formed an EGP Panel 2 'Pool' group composed of all Panel 2 subjects.

SNP genotype data were coded 1, 0, and -1 corresponding to major allele homozygote, heterozygote, and minor allele homozygote. For consistency of the genotype code across populations, major and minor alleles were always classified by the allele frequency in the Pool data. For population-specific data, we calculated MAF within each population. For Panel 1 and pooled Panel 2 data, we calculated MAF using ethnically mixed data. We divided SNPs into two groups: common SNPs where the MAF was $\geq$0.05, and rare SNPs with MAF $<$0.05.

HapMap SNPs were genotyped in four population samples, including 30 CEPH trios, 45 unrelated JPT, 45 unrelated CHB, and 30 YRI trios. HapMap Public Release 20 has genotype information for about 3.7 million SNPs

(~1.2 SNP/kb across the human genome). SNP genotype data were downloaded from the HapMap website. Only the 210 unrelated individuals were included in our analysis. As with the EGP data, we combined CHB and JPT data as 'Asian'. We matched HapMap and EGP SNPs according to reference chromosome positions in dbSNP build 124. If the genotyping orientation was different between EGP and HapMap, HapMap SNP nucleotide data were converted into the complementary nucleotide code.

Composite linkage disequilibrium

Standard measures of LD, including r^2 and D', require assumptions of random mating and Hardy–Weinberg equilibrium (HWE) for phase-unknown data.[15] These assumptions may not be met for EGP Panel 1 data, where ethnic identifiers have been removed from individual samples or for Panel 2 Pool data where ethnic identifiers were ignored. Therefore, instead of r^2, we used in our analysis a measure of composite LD proposed by Weir and Cockerham.[16,17] Composite LD ($\Delta_{AB}=D_{AB}+D_{A/B}$) measures the association of alleles from different loci A and B on the same gamete (gametic LD, D_{AB}), as well as on different gametes (nongametic LD, $D_{A/B}$). $D_{A/B}$ is the usual measure of LD, $D_{AB}=p_{AB}-p_Ap_B$, whereas nongametic LD is $D_{A/B}=p_{A/B}-p_Ap_B$. Where p_{AB} is the frequency of gamete AB, $p_{A/B}$ is the frequency of alleles A and B on two different gametes, p_A and p_A are the frequencies of alleles A and B at two loci. An advantage of the composite LD measure (Δ_{AB}) is that it can be calculated from genotype data directly without requiring an assumption of random mating. In addition, it provides a robust method to test for LD, maintains the correct type I error rate whether or not there is departure from HWE at either locus.[18,19] In the case of random mating, $D_{A/B}=0$, and the composite LD reduces to the usual gametic LD D_{AB}. A test statistic for composite LD proposed by Weir[15]

$$\rho^2_{AB}=\frac{\Delta^2_{AB}}{(p_A(1-p_A)+D_A)(p_B(1-p_B)+D_B)}$$

is based on a normalization of Δ_{AB}. In this expression, $D_A=p_{AA}-p_A$ and $D_B=p_{BB}-p_B$ are the deviations from HWE at each locus, p_{AA} and p_{BB} are the frequencies of genotypes AA and BB. For n individuals, $n\rho^2$ has an approximate $\chi^2_{(1)}$ distribution when $\Delta_{AB}=0$.[15] In most cases, ρ^2 and the gametic LD measure r^2 are very similar.[19] Finally, with our genotype coding, ρ is equivalent to the simple linear correlation coefficient of genotype data at two loci.[20]

Tag SNPs

We employed a greedy algorithm proposed by Carlson *et al*[21] to select tag SNPs from the set of common SNPs for each gene. First, we calculated ρ^2 for all possible pairs of common SNPs within a gene. For each gene, the greedy algorithm selects a SNP where ρ^2 is greater or equal to 0.8 with the largest number of other SNPs, and places these correlated SNPs into one bin. The binning process is iterated for all remaining unbinned SNPs, and continues until ρ^2 is less than 0.8 for all remaining pairs of SNPs. These SNPs are each placed into singleton bins containing only themselves.

We generalized the greedy algorithm to construct a parsimonious set of tag SNPs for multiple populations. As before, we first calculate ρ^2 for all pairs of common SNPs within a genome region separately for each ethnic group. We then execute the following three steps.

1. For each SNP, we count the number of SNPs that have ρ^2 greater or equal to a specified threshold with the SNP. This is done independently for each ethnic group.
2. We sum up the counts for each SNP across ethnic groups. The SNP with the largest sum is selected as a tag SNP.
3. For each ethnic group, we bin SNPs for which ρ^2 exceeds the threshold with the tag SNP.

Steps 1–3 are iterated for all remaining unbinned SNPs within each ethnic group until the only remaining SNPs are those whose sum equals 1. These SNPs are placed into singleton bins containing only themselves.

We note that this algorithm does not require that the different ethnic groups start with the same set of common SNPs. Furthermore, LD patterns may vary between populations so that the set of SNPs binned at each step may differ by ethnic group. We implemented this algorithm into a freely available software *mPopTag* (http://dir.niehs.nih.gov/direb/mpoptag).

For each of the gene regions resequenced by EGP, we used HapMap data to select tag SNPs. We evaluated these tag SNPs against EGP genotype data by calculating the 'gene-tagging proportion', that is, the percent of common EGP SNPs in a gene that are in high LD ($\rho^2\geq 0.8$) with at least one tag SNP. We investigated a simple strategy to increase gene-tagging proportions by supplementing HapMap tag SNPs. For EGP common SNPs that were not in high LD ($\rho^2 < 0.8$) with any HapMap tag, we used the greedy algorithm to construct LD bins. The supplemental tag SNPs were chosen either to tag all bins or only multi-SNP bins.

Simulations

EGP gene resequencing often excluded portions of large introns.[13] HapMap may have SNPs within such unresequenced 'holes' and inclusion of these SNPs might improve HapMap gene-tagging proportions.[22] To estimate the effect of HapMap SNPs in holes on our estimation of gene-tagging proportion, we simulated genes with and without holes using ENCODE data. First, we simulated HapMap SNPs by randomly sampling common SNPs in ENCODE regions at a density comparable to HapMap. To better approximate HapMap SNPs, we restricted sampling to 'RS

SNP' (ie SNPs that were in dbSNP before ENCODE resequencing, http://www.hapmap.org/downloads/encode1.html.en). Second, the contiguous region of resequenced and unresequenced segments of each EGP Panel 2 gene was simulated by randomly placing it within ENCODE regions. For simulated genes both with and without holes, we applied the greedy algorithm to select tag SNPs. We then calculated tagging proportions for common SNPs in the resequenced regions. A similar simulation strategy was used to investigate the effect of HapMap SNPs in flanking regions on gene-tagging proportions.

The small sample size of EGP ethnic groups could lead to biased estimates of gene-tagging proportions. We used the coalescent method implemented in COSI software[23] to simulate genotype data of a 50-kb gene for 10 000 individuals in each of four ethnic groups (European, African-American, African, and Asian). Analogous to HapMap, we randomly sampled 90 individuals for each ethnic group and selected tag SNPs for each ethnic group. Finally, for each ethnic group, we compared tagging proportion estimates for a large sample of 1 000 individuals to tagging proportion estimates for a small sample of 24 individuals (EGP sample size).

We also performed simulations using EGP Panel 2 data to evaluate the effect of a small number of HapMap SNPs that were missing from EGP. We randomly sampled a small subset of EGP common SNPs and added them to the set of HapMap SNPs found in EGP. For both these sets of SNPs, we used EGP genotype data to select tag SNPs. We then calculated gene-tagging proportions in each of the ethnic populations for the two tag SNP sets.

Results

For common SNPs (MAF$\geq$0.05), EGP Panel 1, EGP Panel 2, and ENCODE only have small differences in SNP density (Table 1). On a genome-wide basis, HapMap Release 20 has approximately 45% of the common SNP density found in EGP and ENCODE. We also examined HapMap SNP densities in the specific regions resequenced by EGP, and found that HapMap has a slightly higher SNP density in Panel 1 regions than Panel 2 regions (Table 1). There were 8852 SNPs in EGP Panel 2 that had MAF$\geq$0.05 in one or more of the three ethnic groups within EGP. Of these SNPs, HapMap had genotyped 2710 (31%). Conversely, there were 3073 HapMap SNPs found in EGP Panel 2 resequenced gene regions for the 127 genes that had MAF$\geq$0.05 in one or more of the three ethnic groups within HapMap. Of these SNPs, EGP had genotype information for 2916 (95%).

We selected tag SNPs using HapMap genotype data for 127 genes in EGP Panel 2, and used EGP resequencing data to evaluate their performance in each of the three ethnic groups. Based on HapMap tag SNPs for each gene in each ethnic group, we found that gene-tagging proportions differed by ethnic group. The median gene-tagging proportions were 48, 78, and 72% for African, Asian, and European groups, respectively (Figure 1). We also investigated our decision to pool the ethnically admixed African-American individuals with the YRI individuals into a single 'African' Group. We find that median gene-tagging proportions for African-Americans, YRI, and the pooled 'African' groups only have minimal difference (data not shown).

In general, EGP resequenced the entire genomic sequence for genes whose size was $<$30 kb, whereas resequencing of genes $>$30 kb excluded portions of large introns.[13] Because HapMap has genotype data on SNPs that were in unresequenced regions and thus were not included in our analysis, HapMap-tagging proportions for genes with unresequenced 'holes' may be biased downward.[22] Using ENCODE data, we simulated the effect of unresequenced holes on tagging proportions and found little evidence of bias (Figure 2). We also examined whether inclusion of additional SNPs available within HapMap beyond the 3′ and 5′ flanking regions resequenced by EGP would substantially improve gene-tagging proportions. Simulation results suggested inclusion of an additional 5 kb to both flanking regions provides only modest improvement in gene-tagging proportions (Figure 2). Increasing flanking regions to as much as 20 kb provided very little additional improvement and required many more tag SNPs (data not shown).

Table 1 SNP density

		No. of common (no. of rare) SNPs per kb		
	Region size (Mb)	*African*	*Asian*	*European*
HapMap	3039.69	0.77 (0.43)	0.64 (0.59)	0.69 (0.54)
Encode	5.00	1.80 (2.11)	1.36 (2.57)	1.52 (2.64)
EGP Panel 1	9.02[a]	1.84 (3.95)[b]	1.84 (3.95)[b]	1.84 (3.95)[b]
HapMap in Panel 1 region	9.02[a]	0.91 (1.13)	0.68 (1.38)	0.72 (1.35)
EGP Panel 2	3.09[a]	2.31 (3.36)	1.47 (4.2)	1.61 (4.06)
HapMap in Panel 2 region	3.09[a]	0.72 (0.31)	0.61 (0.44)	0.68 (0.37)

[a]Size of cumulative regions resequenced by EGP excluding unresequenced 'holes' in large introns.
[b]SNP density is based on ethnic-mixed samples and are not ethnic-specific estimates.

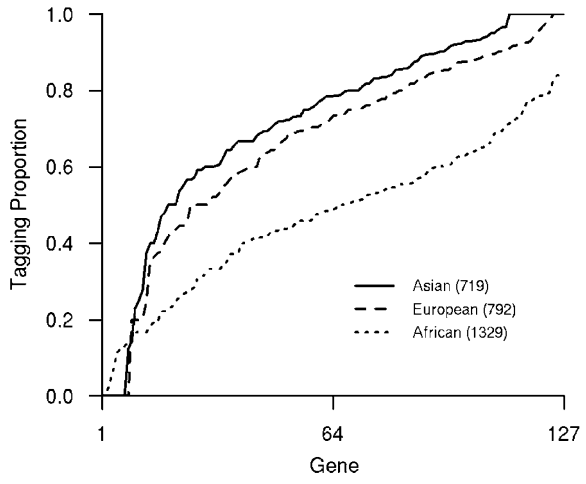

Figure 1 Tagging proportions for individual genes in EGP Panel 2 using tag SNPs based on HapMap genotype data and assessed against EGP resequenced data for each of three ethnic groups. Genes are sorted by tagging proportions independently for each ethnic group. Legend shows number of tag SNPs in parenthesis.

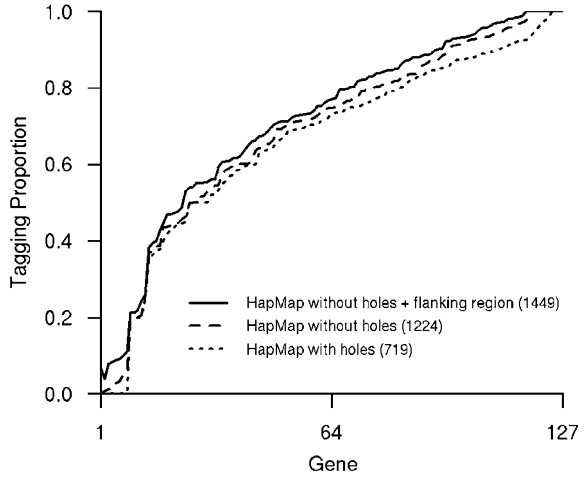

Figure 2 Effect of holes and flanking regions on gene-tagging proportions. Results of 1000 simulations based on ENCODE data (see Materials and methods). Genes are sorted by gene-tagging proportions independently for each ethnic group. The number of tag SNPs are shown in parenthesis.

HapMap SNPs that are not included in EGP could lead to underestimation of the gene-tagging proportions. In total, there were 157 HapMap SNPs that were common in at least one HapMap ethnic group (117, 87, and 105 in African, Asian, and European, respectively), but were not found in EGP. However, the majority of the missed SNPs (72, 67, and 83 in the three ethnic groups, respectively) were in high LD $\rho^2 \geq 0.8$ with another HapMap SNP that did have a match in EGP. The results of 100 simulations suggest that the 157 HapMap SNPs missing from EGP have minimal effect on gene-tagging proportions and, on average, result in a 2% increase in median tagging proportion in the three ethnic groups.

The small sample size of EGP might bias gene-tagging proportion estimates. We used simulations to compare tagging proportions from a sample size of 24 or 1000. The results of 100 simulations suggest that gene-tagging proportion estimates at EGP sample sizes of 24 individuals have minimal bias (data not shown).

We applied the strategy described in Materials and methods for supplementing the set of HapMap tag SNPs. If supplemental tag SNPs for all untagged LD bins are included, then all gene-tagging proportions are increased to 1.0, but this requires a large number of additional tag SNPs, because there are many LD bins with a single SNP. We therefore considered the more efficient strategy of only adding tag SNPs for untagged multi-SNP LD bins. The results in Figure 3 show that this strategy improves the tagging proportions with a modest increase in the number of tag SNPs. For example, in Europeans, an increase from 792 to 962 tag SNPs (1.3 additional tag SNPs/gene) resulted in an increase in the median gene-tagging proportion from 0.72 to 0.87. In contrast, a total of 1537 SNPs would be required to tag all LD bins.

For the 391 genes in EGP Panel 1, ethnic-specific data are not available. To investigate the utility of using ethnic-mixed Panel 1 genotype data to pick tag SNPs, we pooled Panel 2 genotype data and compared the results of the Pool against the ethnic-specific standards. Only 47 (0.3%) of the 16 195 correlated SNP pairs ($\rho^2 \geq 0.8$) in the Pool were not correlated in any ethnic group, suggesting there are minimal false-positive correlations. In addition, 90% of correlated SNP pairs in the Pool were correlated in three or more ethnic groups. Thus, ρ^2 calculated from Pool data appears to reflect LD structure in component populations. The results in Figure 3 show that adding tag SNPs for multi-SNP bins identified from Panel 2 Pool genotype data can improve the gene-tagging proportions of HapMap. For example, in the European sample, an increase from 792 to 1255 tag SNPs (3.6 additional SNPs/gene) resulted in an increase in the median gene-tagging proportion from 0.72 to 0.83. A total of 2228 tag SNPs would be required to include all singleton bins from the Pool and increases median gene-tagging proportion from 0.72 to 0.93.

We applied the generalized greedy algorithm described in Materials and methods to select multipopulation tag SNPs for the three HapMap populations and identified 1674 tag SNPs, of which 959 tagged multi-SNP bins. We evaluated gene-tagging proportions of these 959 tag SNPs in EGP Panel 2 data (Figure 4). The results show that the median gene-tagging proportions were 0.42, 0.74, and 0.74 for African, Asian, and European populations respectively. Median gene-tagging proportions could be increased to 0.55, 0.8, and 0.78, respectively by using all 1674 tag SNPs. Using the supplemental tag SNP strategy described in Materials and methods and applying the multipopulation

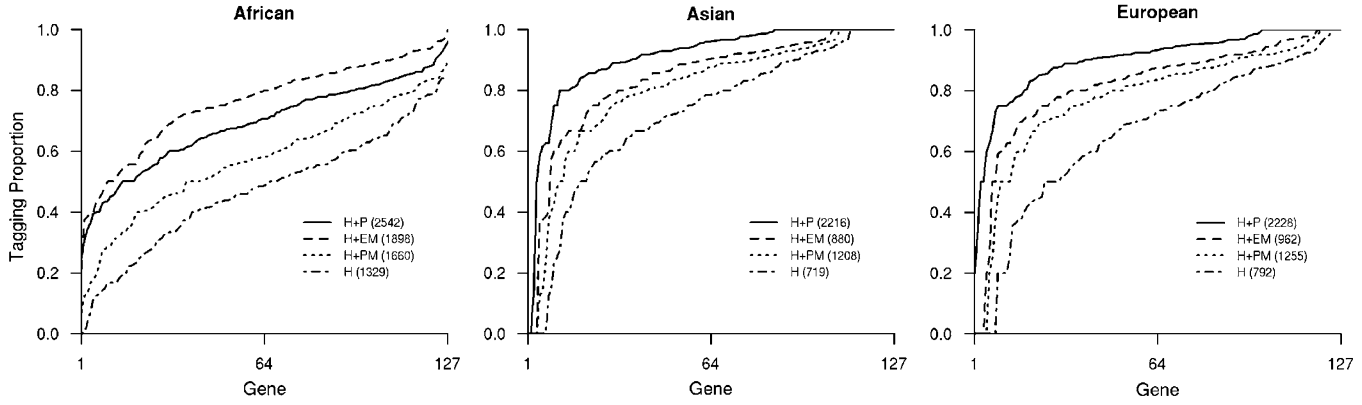

Figure 3 Distribution of gene-tagging proportions for different strategies of augmenting HapMap tag SNPs. For each strategy, genes are sorted by tagging proportion independently for each ethnic group, with number of tag SNPs required shown in parenthesis. H, HapMap tag SNPs; EM, EGP ethnic-specific tag SNPs for multi-SNP LD bins; PM, EGP Pool tag SNPs for multi-SNP bins; P, EGP Pool tag SNPs for all bins (multi-SNP bins + singleton bins) tags.

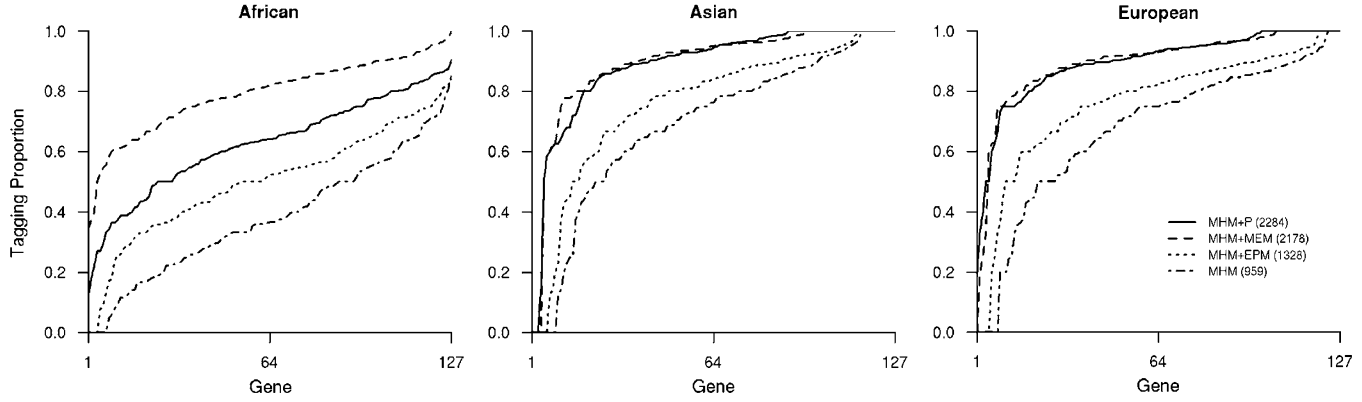

Figure 4 Distribution of gene-tagging proportions for different multipopulation tag SNP strategies. For each strategy, genes are sorted by tagging proportion independently for each ethnic group, with number of tag SNPs required shown in parenthesis. Because they are multipopulation tag SNPs, the number of tag SNPs is the same for all ethnic groups. MHM, multipopulation HapMap tag SNPs for multi-SNP LD bins; MEM, multi-population EGP tag SNPs for multi-SNP LD bins based on ethnic-specific genotype data; PM, EGP Pool tag SNPs for multi-SNP bins; P, EGP Pool tag SNPs for all bins (multi-SNP bins + singleton bins).

tag SNP algorithm to EGP Panel 2 ethnic-specific data, we added 1219 multi-SNP bin tag SNPs (for a total of 2178 tag SNPs), with resulting median gene-tagging proportions of 0.81, 0.94, and 0.92 for the three populations. We also augmented the multi-population HapMap tag SNPs with 369 SNPs from Panel 2 Pool multi-SNP bins and obtained median gene-tagging proportions of 0.52, 0.84, and 0.82 for the three populations. Adding to the multipopulation HapMap tags, all tag SNPs from the Pool (for a total of 2284 tag SNPs), increased tagging proportions to 0.64, 0.93, and 0.93 in the three populations (Figure 4).

For EGP Panel 2 genes, there were on average approximately 3 nsSNPs per gene. The majority of these nsSNPs (~82% in non-African groups and ~72% in the African group) were rare (MAF <0.05). HapMap did not have genotype data on roughly 40% of common and 87% of rare nsSNPs (Table 2). About 30% of the missed common nsSNPs are in high LD with a common SNP in HapMap, but only a very small proportion of rare nsSNPs are in high LD with a common HapMap SNPs. Therefore, approximately

Table 2 Number of nonsynonymous SNPs per gene among EGP and EGP-matched HapMap SNPs

		EGP		*HapMap*	
Panel	*Population*	*Common*	*Rare*	*Common*	*Rare*
EGP Panel 1	European	—	—	0.4	1.15
	Asian	—	—	0.36	1.24
	African	—	—	0.47	1.14
	Pool	0.56	2.35	0.42	1.08
EGP Panel 2	European	0.60	2.79	0.42	0.36
	Asian	0.59	2.80	0.41	0.39
	African	0.95	2.44	0.46	0.31
	Pool	0.77	2.62	0.47	0.27

70% of common nsSNPs and 15% of all (rare plus common) nsSNPs in EGP can be tagged by a common HapMap SNPs. Even if we augment all common HapMap SNPs with all rare nsSNPs in HapMap, only 26% of all nsSNPs in EGP are tagged at $\rho^2 \geq 0.8$. Using this same augmented set of tag SNPs, we found that the multi-marker evaluation method incorporated in Haploview software[24] increased the tagging proportion to 30%.

Discussion

Using ENCODE data, it has been argued that HapMap has adequate SNP density for whole-genome scans. However, HapMap SNP density may pose a problem for some individual candidate genes. ENCODE regions include less than 20 genes and this is an inadequate sample to assess gene-tagging proportions. Using tag SNPs selected from HapMap and applying them to EGP genotype data of 127 genes, we found that tagging proportions were low for nearly half of genes, particularly, when evaluated in African samples.

Our estimation of HapMap-tagging proportions could be biased downward for several reasons. First, EGP did not resequence portions of large introns (holes) and had limited data on flanking regions. We evaluated the possibilities that the inclusion of HapMap SNPs in these regions might improve gene-tagging proportions. Simulations based on ENCODE data suggest that accounting for HapMap SNPs in holes, or in an additional 5 to 20 kb of both 5′ and 3′ flanking sequence, would provide only modest improvements in HapMap gene-tagging proportions for EGP resequenced gene regions. Although inclusion of larger flanking regions might improve gene-tagging proportions, such inclusion might not be cost effective for candidate gene studies. Second, using simulations in EGP Panel 2 data, we evaluated whether the small number of common HapMap SNPs that are missing from EGP affect tagging proportion. Our results suggest that their inclusion would provide minimal improvement in tagging proportions. Finally, we used simulation to investigate the effect of small EGP sample size, but found minimal bias in gene-tagging proportion estimates.

Tagging proportion is a commonly used threshold metric of how well a set of genotyped SNPs captures ungenotyped variants.[1,25,26] However, one must be cautious when using the specified threshold to estimate sample size required for an association study. Because sample size and power to detect a causal variant are not linearly related, merely adjusting sample size requirements by the reciprocal of the threshold is not sufficient to achieve a specified power.[26] The summary metric average maximum r^2 suffers the same problem.[26] For a more complete discussion of this issue and a strategy for obtaining more accurate estimates of sample size, the reader should consult the papers of Jorgenson and Witte.[26,27]

HapMap-tagging proportions can be improved by adding supplemental tag SNPs based on ethnic-specific resequencing data. We noted that gene-tagging proportions in Asians and Europeans can be substantially improved by adding a small number of tag SNPs for multi-SNP bins not yet tagged by HapMap. Gene-tagging proportions can also be improved for Africans but, because of the fine-grained LD structure, require many more tag SNPs.

Despite its lack of individual ethnicity information, EGP Panel 1 represents a rich resource of SNP information that might be useful for tag SNP selection. To examine this possibility, we pooled the EGP Panel 2 genotype data and used these data as a surrogate for EGP Panel 1. This is an appropriate surrogate given that EGP Panel 1 and 2 are similar in the number of people from different ethnic groups, gene function, gene size, and SNP density (http://www.genome.utah.edu/genesnps). EGP Panel 2 Pool data showed that the vast majority of SNP pairs that were correlated in the Pool were also correlated in each of several ethnic groups. We show that tag SNPs from EGP Panel 2 Pool data can augment HapMap tag SNPs to increase gene-tagging proportions, although these tags are not as efficient as tag SNPs from ethnic-specific data. We conclude from these results that the detailed resequencing information on 391 EGP Panel 1 genes may be used to select tag SNPs for multiple populations.

An advantage of multipopulation tag SNPs is that a single set of SNPs can be genotyped in multiple populations, rather than developing different panels of tag SNPs for each population. A disadvantage is that the number of tag SNPs will be larger than the number of tag SNPs in any one ethnic-specific group. Furthermore, the SNPs in an LD bin defined by a multipopulation tag SNP can differ by population, and thus multiple LD or haplotype maps are still needed to analyze the genotype data of multipopulation tag SNPs.

HapMap contained a much higher percentage of rare nsSNPs in EGP Panel 1 gene regions than in EGP Panel 2 gene regions (Table 2). We believe the difference is because Panel 1 data were deposited into dbSNP before HapMap, whereas Panel 2 data were deposited after the creation of HapMap. Thus, Panel 2 data are likely to be representative of the vast majority of genes that have not been extensively resequenced. Although HapMap was not intended to provide coverage for rare SNPs, efforts were made to genotype all known nsSNPs.[1] Similar to results of Barrett *et al*,[11] our results based on EGP Panel 2 data suggest that HapMap provided information for the majority of common nsSNPs, but is of marginal value for the 80% of nsSNPs that are rare. Using a multimarker tag SNP evaluation method provided some improvement in nsSNP-tagging proportion, but the majority of nsSNPs remained untagged.

HapMap is a resource for whole-genome association studies,[1] and is also a powerful resource for other uses,

including the selection of tag SNPs for candidate gene studies. But because HapMap is an incomplete catalog of SNPs, its successful use in candidate gene studies depends on whether this incomplete catalog provides adequate information on the untyped SNPs in genes. Our results suggest HapMap-tagging proportions are low for many genes and that investigators may wish to augment HapMap SNPs with additional SNPs from gene resequencing data. Both EGP Panel 1 and Panel 2 provide a rich SNP resource for a large selection of genes that investigators can use to supplement HapMap tag SNPs. As the cost of resequencing continues to decline, such resources will be available on a larger selection of genes.

Acknowledgements

We thank the anonymous reviewers, whose comments and suggestions greatly improved the manuscript. This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

References

1 The International HapMap Consortium: a haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
2 The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
3 Olden K, Wilson S: Environmental health and genomics: visions and implications. *Nat Rev Genet* 2000; **1**: 149–153.
4 Collins FS, Brooks LD, Chakravarti A: A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 1998; **8**: 1229–1231.
5 Evans DM, Cardon LR: A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet* 2005; **76**: 681–687.
6 Ahmadi KR, Weale ME, Xue ZY *et al*: A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat Genet* 2005; **37**: 84–89.
7 Howie BN, Carlson CS, Rieder MJ, Nickerson DA: Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum Genet* 2006; **120**: 58–68.
8 Ireland J, Carlton VE, Falkowski M *et al*: Large-scale characterization of public database SNPs causing non-synonymous changes in three ethnic groups. *Hum Genet* 2006; **119**: 75–83.
9 Cargill M, Altshuler D, Ireland J *et al*: Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999; **22**: 231–238.
10 Glatt CE, DeYoung JA, Delgado S *et al*: Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nat Genet* 2001; **27**: 435–438.
11 Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659–662.
12 Taylor JA, Xu Z, Kaplan NL, Morris RW: How well do HapMap haplotypes identify Common haplotypes of genes? A comparison with haplotypes of 334 genes resequenced in the Environmental Genome Project. *Cancer Epidemiol Biomarkers Prev* 2006; **15**: 133–137.
13 Livingston RJ, von Niederhausern A, Jegga AG *et al*: Pattern of sequence variation across 213 environmental response genes. *Genome Res* 2004; **14**: 1821–1831.
14 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
15 Weir B: *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates, 1996.
16 Weir BS, Cockerham CC: Complete characterization of disequilibrium at two loci. *Mathematical Evolutionary Theory* 1989.
17 Weir BS: Inferences about linkage disequilibrium. *Biometrics* 1979; **35**: 235–254.
18 Schaid DJ: Linkage disequilibrium testing when linkage phase is unknown. *Genetics* 2004; **166**: 505–512.
19 Weir BS, Hill WG, Cardon LR: Allelic association patterns for a dense SNP map. *Genet Epidemiol* 2004; **27**: 442–450.
20 Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* 2003; **73**: 115–130.
21 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
22 Pe'er I, Chretien YR, de Bakker PI, Barrett JC, Daly MJ, Altshuler DM: Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am J Hum Genet* 2006; **78**: 588–603.
23 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 2005; **15**: 1576–1583.
24 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
25 Zeggini E, Rayner W, Morris AP *et al*: An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat Genet* 2005; **37**: 1320–1322.
26 Jorgenson E, Witte JS: Coverage and power in genomewide association studies. *Am J Hum Genet* 2006; **78**: 884–888.
27 Jorgenson E, Witte JS: A gene-centric approach to genome-wide association studies. *Nat Rev Genet* 2006; **7**: 885–891.