How Well Do HapMap Haplotypes Identify Common Haplotypes of Genes? A Comparison with Haplotypes of 334 Genes Resequenced in the Environmental Genome Project

Jack A. Taylor,[1,2] Zong-Li Xu,[2] Norman L. Kaplan,[3] and Richard W. Morris[4]

[1]Laboratory of Molecular Carcinogenesis, [2]Epidemiology Branch, and [3]Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park; and [4]Department of Anesthesiology, Duke University Medical Center, Durham, North Carolina

Abstract

One of the goals of the International HapMap Project is the identification of common haplotypes in genes. However, HapMap uses an incomplete catalogue of single nucleotide polymorphisms (SNPs) and might miss some common haplotypes. We examined this issue using data from the Environmental Genome Project (EGP) which resequenced 335 genes in 90 people, and thus, has a nearly complete catalogue of gene SNPs. The EGP identified a total of 45,243 SNPs, of which 10,780 were common SNPs (minor allele frequency ≥0.1). Using EGP common SNP genotype data, we identified 1,459 haplotypes with frequency ≥0.05 and we use these as ''benchmark'' haplotypes. HapMap release 16 had genotype information for 1,573 of 10,780 (15%) EGP common SNPs. Using these SNPs, we identified common HapMap haplotypes (frequency ≥0.05) in each of the four HapMap ethnic groups. To compare common HapMap haplotypes to EGP benchmark haplotypes, we collapsed benchmark haplotypes to the set of 1,573 SNPs. Ninety-eight percent of the collapsed benchmark haplotypes could be found as common HapMap haplotypes in one or more of the four HapMap ethnic groups. However, collapsing benchmark haplotypes to the set of SNPs available in HapMap resulted in a loss of haplotype information: 545 of 1,459 (37%) benchmark haplotypes were uniquely identified, and only 25% of genes had all their benchmark haplotypes uniquely identified. We resampled the EGP data to examine the effect of increasing the number of HapMap SNPs to 5 million, and estimate that ~40% of common SNPs in genes will be sampled and that half of the genes will have sufficient SNPs to identify all common haplotypes. This inability to distinguish common haplotypes of genes may result in loss of power when examining haplotype-disease association. (Cancer Epidemiol Biomarkers Prev 2006;15(1):133–7)

Introduction

The human genome contains an estimated 10 million single nucleotide polymorphisms (SNPs), averaging one SNP for every 300 bp (1). Although many SNPs within genes have yet to be identified, an average sized gene of 25 kb can have >80 SNPs (2). This large number of SNPs poses a problem for association studies based on candidate genes because costs limit the number of SNPs that can be genotyped. One strategy to reduce genotyping costs is to select a subset of SNPs that ''tag'' haplotypes of a gene (3). The success of this strategy depends on knowing the common haplotypes within a gene, but an incomplete catalogue of SNPs within a gene may result in an incomplete catalogue of gene haplotypes. In particular, two or more haplotypes that can be distinguished with complete SNP information may appear as a single ''composite'' haplotype when an incomplete set of SNPs is used for haplotype construction. The use of composite haplotypes in candidate gene association studies can result in loss of statistical power if disease is associated with only one member of a composite haplotype.

One of the goals of the International HapMap Project is to identify common haplotypes across the human genome (4). HapMap release 16 has genotyped 269 people from four ethnic groups for ~1 million SNPs. The resulting genotype data is available at the HapMap web site (http://www.hapmap.org) and there are plans to increase the total to 5 million SNPs. Even following the planned expansion, the catalogue of gene SNPs will be incomplete and the degree to which the SNPs in HapMap can discern the common haplotypes of genes is unknown.

One way to ensure the discovery of most common SNPs in genes, and consequently, the most common gene haplotypes, is by resequencing genes from a large number of people. The largest gene resequencing effort thus far is the National Institute of Environmental Health Science Environmental Genome Project (EGP; ref. 5). The large size and ethnic diversity of this sample make it likely that the EGP will find most common SNPs in the four ethnic groups studied by HapMap. In this report, we use EGP resequencing data to assess whether the SNPs in HapMap are sufficient to identify most common haplotypes in genes.

Materials and Methods

The EGP used DNA from 90 unrelated individuals drawn from the Polymorphism Discovery Resource (6), which included 24 African-Americans, 24 Asian-Americans, 24 European-Americans, 12 Hispanic-Americans, and 6 Native Americans, with equal numbers of males and females. As a requirement for making these samples available for scientific use, information on ethnicity was stripped and destroyed from all samples

Received 8/19/05; revised 10/5/05; accepted 11/1/05.

Grant support: Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Cancer Epidemiology Biomarkers and Prevention Online (http://cebp.aacrjournals.org/).

Requests for reprints: Jack A. Taylor, National Institute of Environmental Health Science, MD A3-05, 111 Alexander Drive, P.O. Box 12233, Research Triangle Park, NC 27709. Phone: 919-541-4631; Fax: 919-541-2511. E-mail: taylor@niehs.nih.gov

doi:10.1158/1055-9965.EPI-05-0641

and was not available on an individual basis. The EGP targeted genes for resequencing that are likely to be involved in susceptibility to environmentally associated disease (5). In the EGP, genes <30 kb were sequenced completely, whereas sequencing of genes >30 kb excluded portions of large introns (2). Polymorphisms and genotype information for the 90 individuals in the EGP are deposited into the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/index.html), and are also available at the EGP GeneSNPs web site (http://genome.utah.edu/genesnps). At the time of our analysis, 45,243 biallelic polymorphisms from a total of 335 genes had been deposited into dbSNP Build 124 and assigned reference sequence numbers, including 3,088 insertion/deletion polymorphisms, although for convenience, we refer to all as SNPs. By examining the date of deposit, we were able to determine how many of the 45,243 were novel at the time of their deposit into dbSNP. We classified common SNPs as those with a minor allele frequency (MAF) ≥0.1. The 335 genes are widely distributed across the genome, with at least two genes found on every autosomal chromosome.

We used PHASE version 2.1 (7) with default settings to identify haplotypes. For EGP genes, we restricted our analysis to common SNPs (MAF ≥0.1), and ran PHASE on each of the 334 genes that had two or more common SNPs. We classified those haplotypes with PHASE-generated frequency estimates ≥0.05 as ''benchmark haplotypes.''

We used HapMap release 16 genotype information on a total of 930,384 SNPs that had been genotyped in 269 individuals from all four ethnic groups (http://www.hapmap.org). We used the dbSNP reference sequence number to find a total of 3,400 SNPs that were present in both the EGP and HapMap data sets; 1,573 of these were EGP common SNPs and for these SNPs HapMap haplotypes were constructed with PHASE using the genotype data from: (*a*) 60 unrelated Utah residents with ancestry from northern and western Europe (CEU), (*b*) 60 unrelated Yoruba from Ibadan, Nigeria (YRI), (*c*) 45 unrelated Han Chinese from Beijing, China (HCB), and (*d*) 44 unrelated Japanese from Tokyo, Japan (JPT). We restricted this analysis to 218 genes that had at least two EGP common SNPs that were also found in HapMap.

HapMap contains a subset of the EGP common SNPs. Thus, to compare haplotypes from the two data sets, we ''collapsed'' benchmark haplotypes to the subset of EGP common SNPs found in HapMap. After collapsing, if two or more benchmark haplotypes in a gene were indistinguishable from one another, the collapsed haplotype was designated a ''composite'' haplotype (Fig. 1).

We investigated the effect of increasing HapMap SNP number on benchmark haplotype ascertainment. Starting with the baseline set of 3,400 EGP SNPs found in HapMap release 16, we randomly selected from the 45,243 EGP SNPs until we obtained a specified fold increase of the baseline set of SNPs. We then collapsed benchmark haplotypes to the set of EGP common SNPs sampled in the expanded panel of SNPs, and determined the number of benchmark haplotypes that could be uniquely identified. We repeated this process 100 times and report average values in Fig. 2.

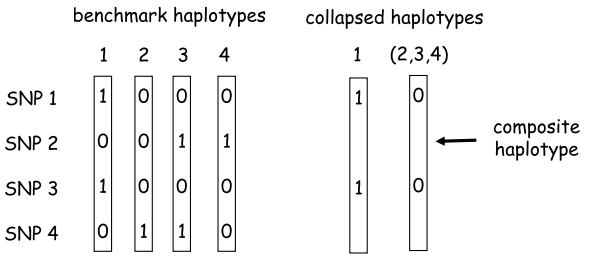

Figure 1. The effect of using a subset of SNPs on benchmark haplotype identification. Consider a gene with four benchmark haplotypes defined by four SNPs, where 0 indicates the major allele and 1 indicates the minor allele. If data are available on SNPs 1 and 3, only two haplotypes will be distinguished. The first haplotype uniquely identifies benchmark haplotype 1, but the second is a ''composite'' of benchmark haplotypes 2, 3, and 4.

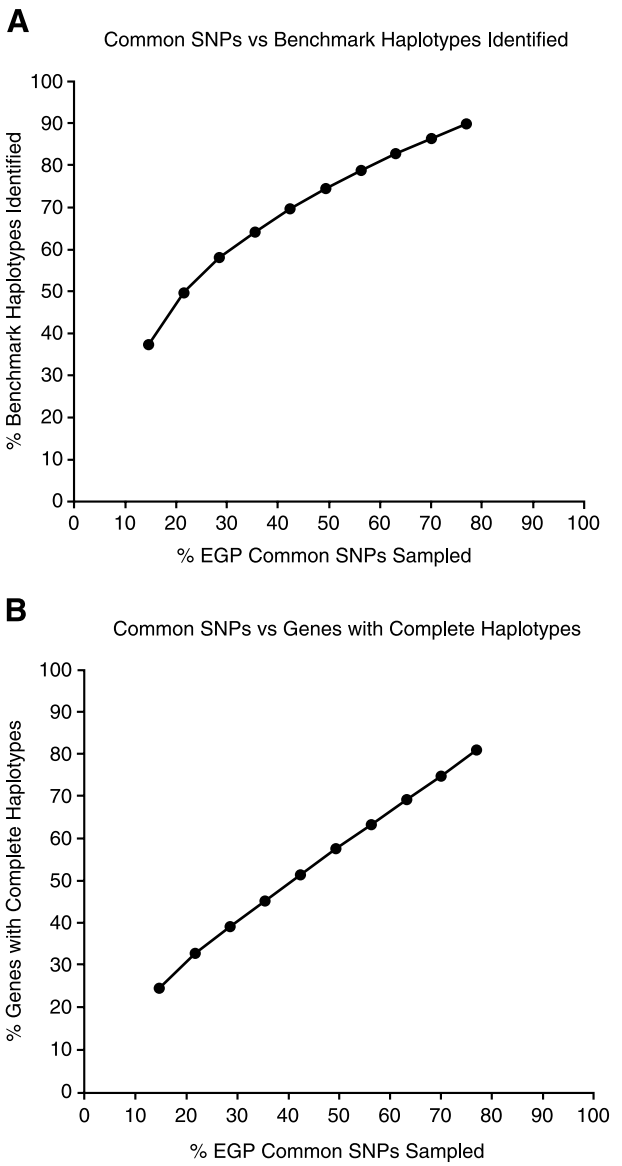

Figure 2. Results simulating increasing numbers of HapMap SNPs. **A,** the relationship between the average proportion of common SNPs sampled and the average proportion of benchmark haplotypes uniquely identified. **B,** the relationship between the average proportion of common SNPs sampled and the average proportion of genes that have all benchmark haplotypes uniquely identified. The first data point is the current percentage of EGP common SNPs sampled using the ~1 million SNPs present in HapMap release 16. Subsequent data points are the average results of 100 simulations for fold increases in the total number of SNPs sampled, up to 10-fold.

Results

Common SNPs and Benchmark Haplotypes in EGP Genes. Reference sequence numbers were available for 45,243 SNPs in 335 EGP genes, 37,439 (83%) of these SNPs

Table 1. Cross-classification of common and rare EGP and HapMap SNPs for four HapMap ethnic groups

EGP	CEU*		HCB		JPT		YRI		Pooled HapMap ethnic groups	
	Common †	Rare †	Common	Rare	Common	Rare	Common	Rare	Common	Rare
Common ‡	1,289	284	1,223	350	1,192	381	1,239	334	1,483	90
Rare ‡	127	1,700	152	1,675	134	1,693	446	1,381	138	1,689
Total	1,416	1,984	1,375	2,025	1,326	2,074	1,685	1,715	1,621	1,779

*HapMap ethnic groups European (CEU), China (HCB), Japan (JPT), Nigeria (YRI).
†HapMap common SNPs with MAF ≥0.1; Rare SNPs with MAF < 0.1.
‡EGP common SNPs with MAF ≥0.1; rare SNPs with MAF <0.1.

were novel in that they had previously not been deposited in dbSNP. Of the 45,243 EGP SNPs, 10,780 were common, with MAF ≥0.1 and 5,867 (54%) of these common SNPs were novel at the time of their deposit in dbSNP. Of the 5,867 EGP common SNPs that were novel deposits, we found that using the most current dbSNP build 124, that 3,873 (36% of 10,780) have yet to be deposited by other groups. Thus, based on EGP resequencing data, we estimate that 36% to 54% of common SNPs are missing from dbSNP.

One EGP gene, *MLP*, had no common SNPs and was removed from the study. Using genotype data for common SNPs, PHASE (7) identified 10,640 haplotypes in the remaining 334 genes. For each gene, PHASE assigned two haplotypes to each of the 90 individuals in the EGP sample. Based on PHASE assignments of haplotypes to individuals, 1,459 of these haplotypes had a frequency of ≥0.05 in the EGP sample and were designated ''benchmark'' haplotypes. Among the 1,459 benchmark haplotypes, 798 (55%) were observed as homozygotes in the EGP sample. For each of the 334 genes, we counted the total number of times that benchmark haplotypes were assigned in the 90 individuals (Supplemental Table S1). The median value per gene was 144 (80%) of 180 assignments and only 54 genes had <90 assignments being benchmark haplotypes. Thus, for most genes, benchmark haplotypes account for the majority of haplotypes assigned by PHASE to the 90 individuals.

HapMap Haplotypes versus Collapsed Benchmark Haplotypes. Using reference sequence numbers, we compared the 45,243 SNPs in EGP with the 930,384 SNPs genotyped in all four ethnic groups in HapMap release 16. We identified 3,400 SNPs that were found in both data sets and 1,573 of these were EGP common SNPs. Thus, HapMap has information on 1,573 of 10,780 (15%) of common SNPs identified by gene resequencing. The classification of matched SNPs into common and rare frequency categories for EGP and HapMap was highly concordant (Table 1). Of the 334 genes, 74 had no matched EGP common SNPs, 42 had one matched EGP common SNP, and 218 had two or more matched EGP common SNPs. We used PHASE to construct haplotypes from HapMap genotype data for the 218 genes with at least two matched SNPs (see Materials and Methods) for each ethnic group and to identify the common HapMap haplotypes (frequency ≥0.05) for each ethnic group (Table 2).

To compare benchmark and HapMap haplotypes, we collapsed benchmark haplotypes to the subset of 1,573 EGP common SNPs that had been matched by HapMap. For the 218 genes with at least two matched SNPs in HapMap, the 944 benchmark haplotypes collapsed into 693 haplotypes. Six hundred and seventy-six of the 693 (98%) collapsed haplotypes matched a common HapMap haplotype in one or more ethnic group (Table 3). Of the 383 common HapMap haplotypes shared across all four ethnic groups, 351 (92%) were found among collapsed benchmark haplotypes (Table 3).

Missing and Composite Haplotypes. Seventy-four genes had no EGP common SNPs included in HapMap. Consequently, the 335 benchmark haplotypes of these genes were designated as ''missing'' from HapMap. For the remaining 260 genes with at least one matched SNP, 1,124 benchmark haplotypes collapsed to 777 haplotypes (Table 4). Five hundred and forty-five of 777 (70%) collapsed haplotypes represented a single benchmark haplotype, whereas the remaining 232 collapsed haplotypes were composites of two or more of the remaining 579 benchmark haplotypes. Eighty-two of 334 genes (25%) had sufficient SNPs included in HapMap to allow all of their benchmark haplotypes to be uniquely identified.

Consequences of Increased SNP Number in HapMap. As HapMap expands to include additional SNPs beyond those in release 16, there will be an increasing number of common SNPs available for haplotype construction. We investigated the consequences of increasing SNP density in HapMap by resampling EGP data and found a monotonic relationship between the proportion of EGP common SNPs sampled and the proportion of benchmark haplotypes that are uniquely identified (Fig. 2A). We estimate that a 5-fold increase in HapMap SNP number to 5 million SNPs would sample 42% of EGP common SNPs and result in 70% of benchmark haplotypes being uniquely identified. Moreover, we found a linear relationship between the proportion of EGP common SNPs sampled and the proportion of genes that have all of their benchmark haplotypes uniquely identified, i.e., have no

Table 2. HapMap common haplotypes in 218 genes with at least two EGP common SNPs matched in HapMap

HapMap ethnic group	Total HapMap haplotypes	Common HapMap haplotypes*
CEU	1,749	840
HCB	1,570	755
JPT	1,473	753
YRI	2,182	940

*Haplotype frequency ≥0.05, estimated by PHASE.

Table 3. Distribution of common HapMap and collapsed benchmark haplotypes among ethnic groups

Number of HapMap ethnic groups in which haplotype was found	Common HapMap haplotypes*	Collapsed benchmark haplotypes matched† (% HapMap haplotypes found)
0	0	17 ‡
1	608	65 (11)
2	226	93 (41)
3	232	167 (72)
4	383	351 (92)

*Common HapMap Haplotypes from each of the four HapMap ethnic groups (from Table 2) categorized by the number of ethnic groups in which they are found.
† Number of collapsed benchmark haplotypes that match a common HapMap haplotype.
‡ Collapsed benchmark haplotypes that had no corresponding common HapMap haplotype.

Table 4. Benchmark and collapsed haplotypes in 334 genes

	Collapsed benchmark haplotypes*			Benchmark haplotypes
	Uniquely identified	Composite (benchmarks represented)	Missing	
218 Genes (≥2 matched SNPs)	507	186 (437)	0	944
42 Genes (1 matched SNP)	38	46 (142)	0	180
74 Genes (0 matched SNPs)	0	0	335	335
Total	545	232 (579)	335	1,459

*Haplotypes formed after collapsing benchmark haplotypes to the subset of EGP common SNPs matched in HapMap. Resulting haplotypes can either uniquely represent a single benchmark haplotype or, in the case of composites, represent two or more benchmark haplotypes. In genes where HapMap had no SNPs matching to EGP common SNPs, all collapsed haplotypes are missing.

missing or composite haplotypes (Fig. 2B). We estimate that a 5-fold increase in HapMap SNP number would result in 51% of genes having all their benchmark haplotypes uniquely identified.

Discussion

The most direct means of finding most common SNPs, and in turn, most common haplotypes, is to resequence DNA from a large number of people. The advantage of gene resequencing for SNP discovery is evident from the EGP in which 83% of all SNPs and 54% of common SNPs had not been previously deposited into dbSNP. Unfortunately, gene resequencing is expensive. A less costly strategy for finding common haplotypes, adopted by the International HapMap Project, uses a subset of SNPs from dbSNP (4). The incomplete catalogue of SNPs available in dbSNP may limit the identification of haplotypes within HapMap. Without resequencing data for comparison, it is unclear how completely HapMap, at its current or projected SNP density, could discern the common haplotypes of genes.

The large number of people and genes resequenced by the EGP, along with the ethnic diversity of the people included in the EGP, helps to ensure that most common SNPs are sampled. However, the lack of specific ethnic information for the 90 individuals genotyped by EGP poses a potential problem for the statistical inference of haplotypes. Like other haplotype construction programs, PHASE assumes that the sample comes from a randomly mating population (7, 8). Consequently, the constructed nature of the EGP sample, along with missing ethnic information, might lead to some anomalies in haplotype inference. However, earlier studies have successfully used PHASE when analyzing ethnically mixed samples (2), suggesting that PHASE may be robust with regard to these problems. Furthermore, in our analysis, we found that more than half of the benchmark haplotypes inferred by PHASE occurred as homozygotes in at least one individual, and so these haplotypes are unambiguously observed within the EGP sample. Finally, we find that 98% of the collapsed benchmark haplotypes from EGP can be found as PHASE haplotypes in at least one of the four populations genotyped by HapMap. Thus, there is independent confirmation for virtually all of the collapsed benchmark haplotypes.

Benchmark haplotypes were constructed using all EGP common SNPs found via resequencing. We used these to examine whether SNP subsets, in particular, the 15% subset of EGP common SNPs included in HapMap release 16, were sufficient to uniquely identify the benchmark haplotypes found within each gene. Our finding that 545 of 1,459 (37%) benchmark haplotypes could be uniquely identified suggests that HapMap release 16 has insufficient SNPs for discerning most common haplotypes in genes. We estimate that 75% of genes do not have complete ascertainment of all their benchmark haplotypes. Missing and composite gene haplotypes can decrease the power of association studies, with the magnitude of the reduction dependent on a combination of factors including the magnitude of risk, frequency of the haplotypes forming the composite haplotype, and sample size.

We also examined the effect of the planned expansion of HapMap which, using dbSNP, will attempt to develop genotype assays for a total of 5 million SNPs. The EGP resequencing found that 54% of EGP common SNPs were missing from dbSNP at the time they were deposited. Recent deposits into dbSNP by Hinds et al. (9) and others have reduced this proportion so that now only 36% of EGP common SNPs have been solely deposited by EGP in dbSNP build 124. These results suggest that ~35% of common SNPs in genes may be missing from dbSNP.

SNP assay conversion rates, i.e., the fraction of SNPs that could be reliably genotyped, range between 60% and 90% for current technologies, and average ~65% for randomly selected SNPs (10). If we assume that 65% of common SNPs are deposited in dbSNP and that HapMap attempts to genotype all 10 million SNPs in dbSNP with an assay conversation rate of 65%, then HapMap should eventually have genotype information on 42% (65% × 65%) of common SNPs in genes. Interestingly, simulation of a 5-fold expansion of HapMap based on resampling of EGP data also produced an estimate that ~40% of common SNPs would be sampled. If we assume that ~40% of common SNPs are genotyped after the HapMap expansion; the results in Fig. 2 suggest that a third of the common haplotypes of genes may not be uniquely identified, and that about half of the genes will have incomplete ascertainment of common haplotypes.

Because not all SNPs can be converted into reliable genotyping assays, the problem of missing and composite haplotypes might persist even if dbSNP had a complete catalogue of common SNPs. The most effective means of reducing the number of missing and composite haplotypes would be to improve the assay conversion rate, particularly for common SNPs in genes. Assay conversion rates exceeding 90% can be obtained for targeted SNPs by designing assays for both DNA strands, by using multiple genotyping technologies, and by other means (10). Unfortunately, only a small proportion of the SNPs deposited into dbSNP have frequency estimates (2), and these estimates are sample-dependent, so that *a priori* identification of common SNPs is problematic.

Although genome resequencing of large numbers of individuals will eventually become available, HapMap currently provides the largest systematic study of SNP variation in the human population. Unlike the relatively small number of genes selected for resequencing in the EGP, HapMap has information on a much larger genomic scale and thus can be used as a resource for virtually any candidate gene and for whole genome association studies. In addition, unlike EGP data in which ethnic identifiers are unknown, HapMap provides data for each of four ethnic groups and includes family trio data for two of the four

groups. Finally, unlike dbSNP, the SNPs within HapMap have been validated in high-throughput genotyping assays so that they can more easily and confidently be applied in studies of disease association.

HapMap is a rich resource for gene haplotype identification when carrying out candidate gene association studies. However, epidemiologists need to be aware that at current and planned levels, HapMap may not identify all common haplotypes of genes and that the loss of information may decrease power to detect disease association.

References

1. Kruglyak L, Nickerson D. A. Variation is the spice of life. Nat Genet 2001; 27:234–6.
2. Livingston RJ, von Niederhausern A, Jegga AG, et al. Pattern of sequence variation across 213 environmental response genes. Genome Res 2004;14: 1821–31.
3. Johnson GC, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233–7.
4. The International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.
5. Olden K, Wilson S. Environmental health and genomics: visions and implications. Nat Rev Genet 2000;1:149–53.
6. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998;8: 1229–31.
7. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68:978–89.
8. Lin S, Cutler DJ, Zwick ME, Chakravarti A. Haplotype inference in random population samples. Am J Hum Genet 2002;71:1129–37.
9. Hinds DA, Stuve LL, Nilsen GB, et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005;307:1072–9.
10. Fan JB, Oliphant A, Shen R, et al. Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 2003;68:69–78.

How Well Do HapMap Haplotypes Identify Common Haplotypes of Genes? A Comparison with Haplotypes of 334 Genes Resequenced in the Environmental Genome Project

Jack A. Taylor, Zong-Li Xu, Norman L. Kaplan, et al.

Cancer Epidemiol Biomarkers Prev 2006;15:133-137.