Screening for Recently Selected Alleles by Analysis of Human Haplotype Similarity

Neil A. Hanchard,[1,2] Kirk A. Rockett,[1,2] Chris Spencer,[3] Graham Coop,[3] Margaret Pinder,[4] Muminatou Jallow,[4] Martin Kimber,[1,5] Gil McVean,[3] Richard Mott,[1] and Dominic P. Kwiatkowski[1,2]

[1]Wellcome Trust Centre for Human Genetics and Departments of [2]Paediatrics and [3]Statistics, University of Oxford, Oxford; [4]Medical Research Council (MRC) Laboratories, Banjul, The Gambia; and [5]Tessella, Abingdon, United Kingdom

There is growing interest in the use of haplotype-based methods for detecting recent selection. Here, we describe a method that uses a sliding window to estimate similarity among the haplotypes associated with any given single-nucleotide polymorphism (SNP) allele. We used simulations of natural selection to provide estimates of the empirical power of the method to detect recently selected alleles and found it to be comparable in power to the popular long-range haplotype test and more powerful than methods based on nucleotide diversity. We then applied the method to a recently selected allele—the sickle mutation at the *HBB* locus—and found it to have a signal of selection that was significantly stronger than that of simulated models both with and without strong selection. Using this method, we also evaluated >4,000 SNPs on chromosome 20, indicating the applicability of the method to regional data sets.

The identification of human alleles that have undergone recent positive selection may yield important insights into the genetic basis of common diseases, and haplotype analysis has been proposed (Sabeti et al. 2002; Toomajian et al. 2003) as a more powerful way of detecting such alleles than conventional approaches based on nucleotide diversity (Tajima 1989; Fu and Li 1993; Fay and Wu 2000). Under a neutral model, most alleles of high population frequency are relatively ancient and are associated with many different haplotypes as a result of recombination. However, when a new allele first arises in the population, it is associated with a single haplotype; if it then undergoes a rapid increase in frequency because of positive selection, there is little time for recombination to break down the ancestral haplotype. Thus, the underlying principle behind haplotype-based approaches to detecting recent selection is that recently selected alleles tend to show lower haplotypic diversity than do ancient alleles of similar frequency. This has been shown to be the case at several loci with previous evidence of recent selection—at the *G6PD* (Sabeti et al. 2002), *LCT* (Bersaglieri et al. 2004), and *HBE* (Ohashi et al. 2004) loci by use of the method of Sabeti et al. (2002) and at the *HFE* locus by use of the method of Toomajian et al. (2003). Although the application and precise interpretation of haplotype-based methods remain unclear, the underlying premise is a promising avenue of investigation.

In this study, we consider an elementary implementation of the haplotype-based approach. Imagine a set of polymorphisms that have been typed in a representative population sample, from which haplotypes have been inferred across a defined genomic region. From this, we can calculate the similarity of the haplotypes (which we call the "haplosimilarity [HS] score") associated with the minor allele of the first polymorphism ("allele X") by passing a sliding window across the region,

$$\mathrm{HS} = \frac{\sum_{t=1}^{T}\sum_{i}^{k} f_{it}^2}{T},$$

where T is the total number of windows across the region, k is the total number of distinct haplotypes associated with allele X within a given window, and f_{it} is the frequency of the ith distinct haplotype associated with allele X in window t. The value of HS ranges from a minimum given by $1/k_{\max}$, where $k_{\max}$ is the maximum possible number of distinct haplotypes for a given win-

Received July 1, 2005; accepted for publication October 17, 2005; electronically published November 16, 2005.

Address for correspondence and reprints: Dr. Neil A. Hanchard, Tropical Metabolism Research Unit, Tropical Medicine Research Institute, University of the West Indies, Kingston 7, Jamaica. E-mail: neil.hanchard@green-oxford.com

Am. J. Hum. Genet. 2006;78:153–159. © 2005 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7801-0017$15.00

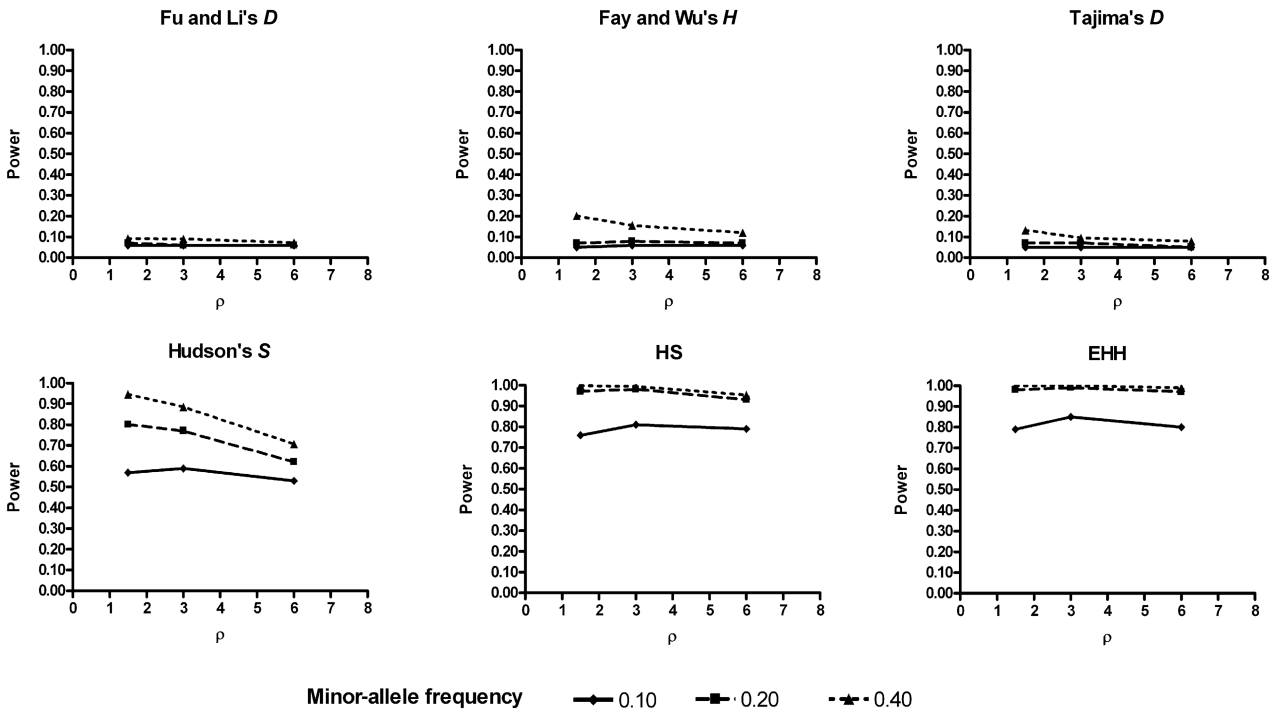

Figure 1 Empirical estimates of power to detect the signature of recent selection. Power is expressed as the proportion of HS scores ($N = 5{,}000$), derived for a simulated SNP undergoing a selective sweep, that exceeded the 95th percentile of scores obtained by simulating the same SNP under neutral conditions. The graphs show the variation in power in relation to recombination rate ($\rho = 1.5$, 3, and 6) and minor-allele frequency. Each line contains data points for one of the three simulated minor-allele frequencies: 0.10 (*blackened squares*), 0.20 (*blackened diamonds*), and 0.40 (*blackened triangles*). In all simulations, nHS scores were calculated using an overlapping sliding-window size of 10 SNPs (see authors' supplementary-data Web site for details).

dow size (i.e., the haplotypes associated with allele X are extremely diverse), to 1 (i.e., all haplotypes associated with allele X are exactly the same). The full set of haplotypes can then be re-sorted by the minor allele of the next SNP, which becomes allele X, and the process is repeated until all SNPs are tested.

To get a preliminary indication of the power of the HS statistic to distinguish a selected from a nonselected allele, we performed coalescent simulations of a genomic region comprising 50 SNPs in a sample of 200 chromosomes, using SelSim (Spencer and Coop 2004). For each set of simulations, the minor allele of the central SNP (SNP 25) was first simulated to attain three different allele frequencies (0.10, 0.20, and 0.40) by the random process of genetic drift. Each set of simulations was run 5,000 times, which created a distribution of HS scores for SNP 25 at each allele frequency. From this, the 95th percentile score was noted. Next, the same allele was simulated to attain the same three frequencies, but this time by undergoing a partial selective sweep (with selection coefficient $s = 500$). Power was calculated as the proportion of scores for the selected SNP that exceeded the 95th percentile of scores for the nonselected (neutral) SNP.

Each of these simulations was performed using three different (but uniform) recombination rates across the region. Taking the recombination parameter $\rho = 4N_e r = 0.4$/kb, with effective population size $N_e = 10{,}000$ and per-generation crossover rate $r = 0.00001$ M/kb (~1 cM/Mb), we simulated the 50 SNPs to be spaced ~1 SNP every 4, 8, and 16 kb, which gave approximations for ρ of 1.6, 3.2, and 6.4, respectively, between SNPs (we approximated these values to 1.5, 3, and 6).

As shown in figure 1, the HS statistic distinguished the selected from the nonselected alleles on >90% of occasions when the allele frequency was $\geq$0.2, irrespective of recombination rate. This was considerably more sensitive than measures of selection based on the nucleotide diversity of the entire population of haplotypes, such as Tajima's *D* test, Fu and Li's *D* test, and Fay and Wu's *H* test (Tajima 1989; Fu and Li 1993; Fay and Wu 2000). The haplotype-partition method developed by Hudson et al. (1994) performed better than the nucleotide-based methods but was less powerful than the HS statistic. For comparison, we performed similar power calculations for the extended haplotype homozygosity (EHH) statistic of the long-range haplotype test (LRH test) and found that the HS statistic performed just as well as the EHH

statistic. EHH is the probability that any two haplotypes extending outward from a SNP will be the same at a given distance away from the SNP. EHH values are therefore directional and can be reported in both 3′ and 5′ directions. We calculated EHH values at the two points farthest away from the SNP being tested and took the arithmetic mean of the two values.

We also used coalescent simulations to test the impact of population demography on our statistic. To do so, the coalescent was simulated using ms (Hudson 2002) with a sample of 200 chromosomes, each with 25 SNPs (to reduce computer intensity), and a uniform recombination rate of $\rho = 3$ between SNPs. A rapid population expansion was simulated using a population that was at a constant size of 10,000 until 500 generations ago, when it expanded by a factor of 100. A population bottleneck was simulated using a population of starting size of 10,000 that underwent a bottleneck 500 generations ago, after which time there was no growth until it recently expanded by a factor of 100 to attain its present size of 10,000. Population structure was simulated using two equal-sized populations of size $N_0/2$ (where N_0 is the effective population size at time 0) that exchanged migrants with probability 0.10. Unlike SelSim, ms does not have the ability to condition on a particular marker; thus, 500–1,000 simulations were run for each model, and a bin of scores for markers with minor-allele frequencies in the range 0.07–0.13 (inclusive) was created. This bin contained 1,939 scores for population growth, 1,300 for bottleneck, and 3,500 for structure, and these were compared with 2,000 values derived from the neutral distribution of a SNP with minor-allele frequency of 10%, simulated using SelSim under similar conditions (200 chromosomes with 25 SNPs each; $\rho = 3$ between SNPs). The proportion of simulated scores that exceeded the 95th percentile of scores derived under a neutral model was 0.18 for a model of population growth, 0.26 for a recent population bottleneck, and 0.07 for population structure, suggesting that the method is somewhat resistant to false-positive confounders resulting from simple demographic models.

Practical application of any haplotype-based test for selection is complicated by the fact that recombination and mutation rates vary across the genome, which may cause variations in haplotype diversity that have nothing to do with selection. For a first approximation, we can control for this by comparing the HS score of a minor SNP allele with that of its major allele—in a manner similar to the relative EHH statistic of the LRH test—since both alleles are subject to exactly the same regional conditions. A second complication is that the haplotype diversity associated with a SNP allele is partly a function of its allele frequency and the allele frequency of surrounding SNPs. To some extent, we can correct for this as follows. Imagine a minor SNP allele that is found on m chromosomes within our sample; we calculate its HS score directly. From the remaining chromosomes that carry the major allele, we randomly select m chromosomes and derive the HS score. If this is repeated on multiple occasions (we performed 1,000 repetitions), it gives a mean and variance estimate of the HS score for the major allele, corrected for the frequency of the minor allele. We then calculate a normalized HS score (nHS),

$$\mathrm{nHS} = \frac{\mathrm{HS_p} - \mathrm{HSq}}{\mathrm{SD_q}} ,$$

where $\mathrm{HS_p}$ is the HS score of the minor allele, $\mathrm{HS_q}$ is the mean of the iterated HS scores for the major allele, and $\mathrm{SD_q}$ is the SD of the HS scores for the major allele. Thus, the diversity of haplotypes carrying the minor SNP allele is expressed in relation to the diversity of haplotypes carrying the major allele.

As a practical test of this approach, we applied the HS method to the sickle hemoglobin polymorphism (HbS) in the β-globin gene (*HBB* [MIM 141900]), using a Gambian population sample. Gambian children who are heterozygous for the HbS allele have a 10-fold reduction in risk of life-threatening complications of malaria (Hill et al. 1991); since most HbS alleles exist in this heterozygous form, this has the effect of selectively increasing the frequency of the allele over several generations. There is also evidence that the HbS allele originated in this population relatively recently (Currat et al. 2002).

For this evaluation, 20 SNPs with minor-allele frequency >10% were chosen from a 414-kb region centered on the HbS mutation (table 1). SNPs were genotyped in 32 Gambian family trios with malaria-affected offspring and 192 unrelated Gambian cord-blood samples. Haplotypes from the family trios were constructed using PHAMILY (Ackerman et al. 2003) and PHASE (Stephens et al. 2001), and the untransmitted parental haplotypes were then used as phase-known haplotypes in the ultimate PHASE reconstruction of unrelated haplotypes. The final available 383 chromosomes were used to derive HS and nHS scores for each of the 20 SNPs.

The HbS allele had an HS score of 0.587 (see table 1) and an nHS score of 56 (fig. 2), both considerably higher than those obtained for any other SNP in the set. To estimate the significance of this result, we performed 1,000 coalescent simulations of a similar-sized region with the same marker spacing and density, first for a neutral SNP and then for a strongly selected ($s = 500$)

Table 1

HS Scores for SNPs in the β-Globin Gene Region

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

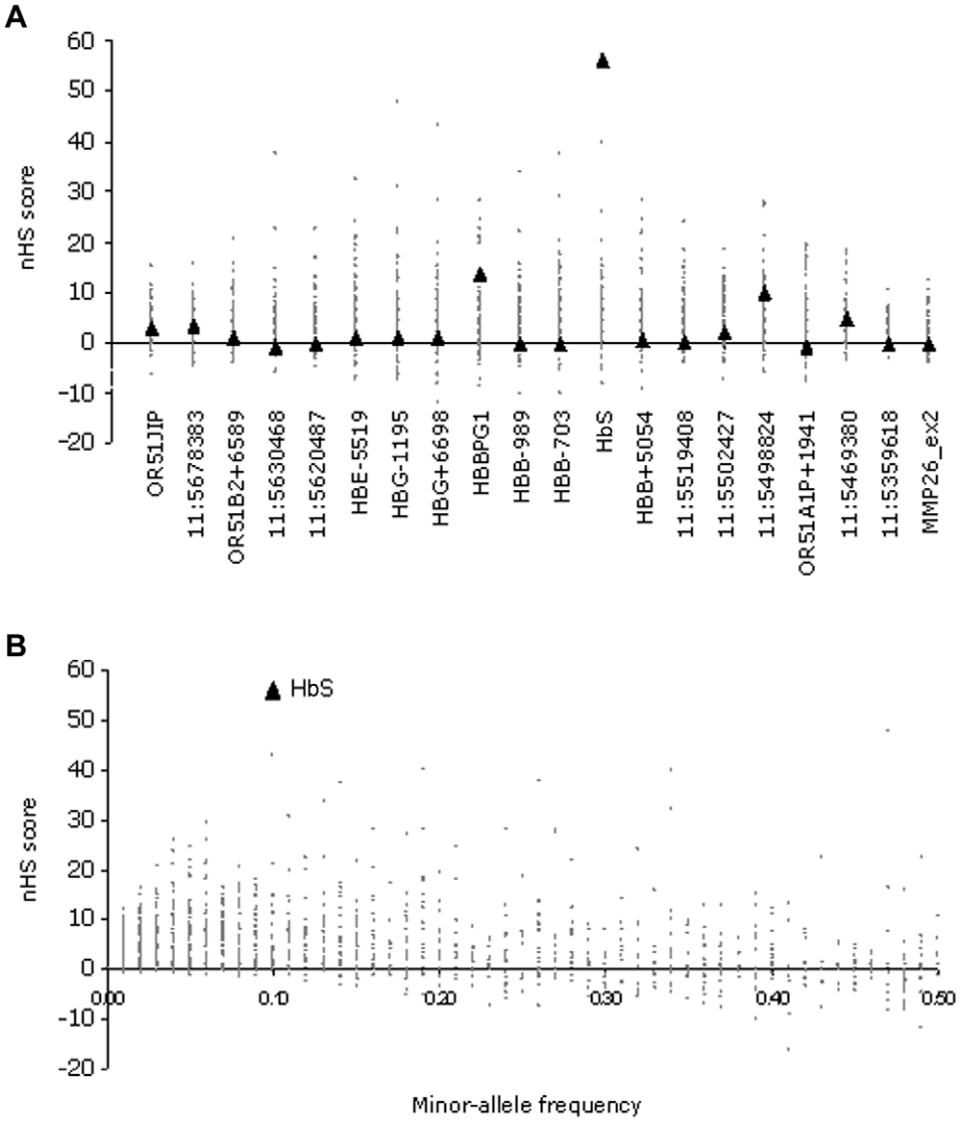

Figure 2 nHS scores in the β-globin gene region. *A*, nHS scores for each of the 20 SNPs are plotted (*blackened triangles*) against SNP position; SNPs are arranged 5′→3′ (*left to right*). *A* and *B*, Scores in gray represent 100 simulations of the region (done for illustrative purposes) that are based on the same number and positioning of SNPs, with a mean recombination rate of 1 cM/Mb across the region. nHS scores were calculated for each of the 2,000 SNPs; they are superimposed on the scores for the *HBB* region in panel A and are displayed by allele frequency, along with the score for HbS (*blackened triangle*), in panel B.

SNP of the same frequency and in the same position as HbS. The nHS score observed for HbS was significantly higher than that obtained for either simulated SNP (with and without selection $P < .001$), which implies that the HbS allele has undergone extremely strong selection in the recent past.

To further evaluate the properties of the nHS score when applied to the kind of data sets that commonly exist—SNP-based haplotypes in which there is no a priori expectation of selection—we applied the HS method to a high-density SNP map across a 10-Mb region of chromosome 20 that has been typed for 4,228 SNPs (Ke et al. 2004). Haplotypes were constructed across regions of ~400 kb by use of the PHASE algorithm (Stephens et al. 2001). Figure 3 summarizes the nHS scores for each of the 4,228 SNPs (full data are available at the authors' supplementary-data Web site; see also MARKER Web site). The results illustrate the potential ability of the nHS statistic to account for differences in allele frequency; most markers had nHS scores <10 (mean ± SD = 0.72 ± 3.8; see fig. 4), and this distribution was similar to that observed for the simulation of the *HBB* region (fig. 2) as well as other in-house data sets (data not shown).

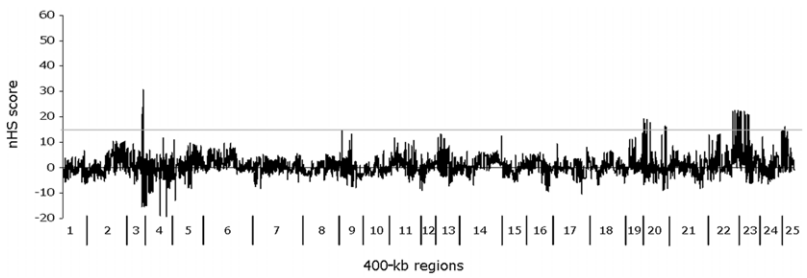

Figure 3 nHS scores for SNPs across 10 Mb of chromosome 20. Each spike represents the nHS score of a SNP. SNPs are arranged 5′→3′ (*left to right*), with a demarcation on the *X*-axis for each of the 25 400-kb regions. Within each region, SNP density ranged from 1 SNP per 1.33 kb to 1 SNP per 4.49 kb. The gray line is an arbitrary cutoff value of 15 based on the distribution of scores (see fig. 4).

Twenty-six SNPs appeared to be outliers in the empirical distribution of scores (i.e., had nHS score >15). To ensure that these values were not simply the result of the relative position of these SNPs along the 400-kb haplotype (in the middle vs. at the end) or the result of our designation of regions, we realigned three regions, placing SNPs with high nHS scores (>15) at different positions along the haplotype (fig. 5). The retested SNPs again yielded scores >15, and the scores of surrounding SNPs again clustered below this value.

Without prior knowledge about any of the SNPs, it is difficult to know whether our observed outliers are true examples of recent selection or simply false positives. The relative rarity of such markers, however, lends itself to their independent evaluation. For example, in region 23, there were 18 such SNPs, all with allele frequency of 0.38 despite being at different positions across the 400-kb region. This is the kind of pattern that might be seen for a number of SNPs hitchhiking on the background of a selected SNP.

Using coalescent simulations, we have provided an empirical test of the power of the HS method to detect the signature of recent selection, and we found it to have close to 80% power to detect reasonably strong selective sweeps, even with a high recombination rate and minor-allele frequencies <0.20. This power is much greater than that of nucleotide-based methods such as Tajima's *D* test and is comparable to that of the LRH test.

Both the LRH test and the HS method assess the similarity of long-range haplotypes, and their similarity in power is therefore not surprising; however, the use of an iterative sliding window and the added resampling step make the HS method particularly suited to evaluating SNPs within a defined region. The sliding-window approach allows the HS method to return a single value for each SNP (rather than values referenced to directional movement or distance away from a SNP) and provides the potential to sum over genotyping errors or phasing inaccuracies that can create the illusion of a unique or dissimilar haplotype. Assessment of the similarity of haplotypes in increments is particularly suited to regions with high SNP density, where almost all haplotypes will be unique. The resampling step (reflected in the nHS score) allows the screening of all SNPs within a circumscribed region without the need to condition on allele frequency; the majority of scores for both simulated and genomic SNPs were clustered below 10, despite inherent differences in recombination rate, SNP density, spacing, and minor-allele frequency.

There are, however, some caveats to the implementation of the statistic. The HS method uses both a sliding-window approach and an iterative resampling step on inferred haplotypes. Therefore, the accuracy of the method is partly dependent on the integrity of the inferred haplotypes, which may be difficult to assess. In addition, although implicit in the examples given, it should be emphasized that the HS method relies on long-distance haplotypes (over several hundred kb) because, over shorter distances, differences in haplotype variability will not be apparent. Also, as implemented, the HS method considers only the minor allele of each SNP. The method can therefore evaluate only alleles at <50% frequency and, critically, assumes the allele to be the derived, as opposed to the ancestral, allele. Distinguishing between these two states is pivotal to the expected HS score of a given allele (which likely explains the very negative scores obtained in regions 3 and 4 on chromosome 20) and should be considered in the interpretation of scores. A further consideration arises in trying to apply the HS method to each SNP across a discrete

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 4 Distribution of nHS scores across chromosome 20

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 5 nHS values for realigned chromosome 20 regions. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

region; some SNPs will be in the middle of the region, whereas others will be on the edge. This could potentially create a bias toward higher scores for some SNPs. Although the neutral simulation of the *HBB* region (fig. 2*B*) and the chromosome 20 data, as well as our preliminary evaluation (data not shown), suggest that the effect of SNP position on the test is minimal, we acknowledge that this is an area that will require further work.

Finally, the HS method is not immune to some of the limitations inherent in most other haplotype-based methods. For instance, it is conceivable that long-range haplotypes could be observed with markers undergoing negative selection (Nordborg and Tavare 2002), although the expectation is that observing such alleles at high frequency is somewhat unlikely. Similarly, although it appears as though the nHS statistic is not likely to misclassify simple demographic processes as resulting from selection, complex demographic interactions can produce significant interaction between distant loci (Thompson and Neel 1997). Such interactions between loci clearly occur (their commonality remains unknown), and, although they are more often seen with low-frequency mutations, definitively distinguishing the genomic signature of selection from that of complex population demography remains difficult. Along with issues of ascertainment bias—inherent in the use of genotype rather than sequence-based data sets—accounting for population demography is a key area of research for the future of haplotype-based applications. The possibility of false positives (via demography or negative selection) and, especially with low-frequency alleles, false negatives (in all of our empirical estimates, power was reduced with decreasing allele frequency) suggests that labels of selection (or nonselection) should not be assigned without considering such attendant confounders.

In light of these considerations, we suggest that the most useful implementation of the HS method is as a screen for selection rather than as a confirmation of it. nHS scores should first be interpreted in the context of the "background scores" provided by surrounding SNPs, and high nHS scores, in particular, should be evaluated on their own merits. In the future, it should be possible to broaden the applicability of the method—for example, to include information on the ancestral state of the minor allele and to better account for ascertainment bias. In the context presented here, however, the HS method is a useful screening tool for the kind of data sets that are fast becoming commonplace in the genetics workplace, and it further advances the idea that the use of haplotypes to detect recent selection can be both sensitive and powerful.

Acknowledgments

N.A.H. was sponsored through a scholarship from the Rhodes Trust, Oxford. This work was financially supported by the MRC, United Kingdom. Permission for the collection, storage, and use of these samples for malaria research (family trios) and genetic research (cord blood) was granted by the Joint Gambian Government/MRC Ethics Committee.

Web Resources

The URLs for data presented herein are as follows:

Authors' supplementary-data Web site, http://www.gmap.net/pub/003 (for supplementary data, including information on operating characteristics)
Ensembl Genome Browser, http://www.ensembl.org/
MARKER Web site, http://www.gmap.net/marker (for public access to haplosimilarity algorithm)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *HBB*)

References

Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, Taylor T, Ward R, Molyneux M, Pinder M, Kwiatkowski DP (2003) Haplotypic analysis of the *TNF* locus by association efficiency and entropy. Genome Biol 4:R24

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Langaney A, Excoffier L (2002) Molecular analysis of the β-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the β^S Senegal mutation. Am J Hum Genet 70:207–223

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM (1991) Common West African HLA antigens are associated with protection from severe malaria. Nature 352:595–600

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster.* Genetics 136:1329–1340

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577–588

Nordborg M, Tavare S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am J Hum Genet 74:1198–1208

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Spencer CC, Coop G (2004) SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20:3673–3675

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Thompson EA, Neel JV (1997) Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am J Hum Genet 60:197–204

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287–297