

## ARTICLE

# A powerful score test to detect positive selection in genome-wide scans

Ming Zhong<sup>1</sup>, Kenneth Lange<sup>2</sup>, Jeanette C Papp<sup>2</sup> and Ruzong Fan<sup>\*,1,3</sup>

One of the surest signatures of recent positive selection is a local elevation of advantageous allele frequency and linkage disequilibrium (LD). We proposed to detect such hitchhiking effects by using extended stretches of homozygosity as a surrogate indicator of recent positive selection. An extended haplotype-based homozygosity score test (EHHST) was developed to detect excess homozygosity. The EHHST conditioned on existing LD and it tested the haplotype version of the Hardy–Weinberg equilibrium. Compared with existing popular tests, which usually lack clear distribution, the EHHST is asymptotically normal, which makes analysis and applications easier. In particular, the EHHST facilitates the computation of an asymptotic  $P$ -value instead of an empirical  $P$ -value, using simulations. We evaluated by simulation that the EHHST led to appropriate false-positive rates, and it had higher or similar power as the existing popular methods. The method was applied to HapMap Phase II data. We were able to replicate previous findings of strong positive selection in 17 autosome genomic regions out of 20 reported candidates. On the basis of high EHHST values and population differentiations, we identified 15 new candidate regions that could undergo recent selection.

*European Journal of Human Genetics* (2010) 18, 1148–1159; doi:10.1038/ejhg.2010.60; published online 12 May 2010

**Keywords:** extended haplotype homozygosity; linkage disequilibrium; positive selection

## INTRODUCTION

In recent years tremendous interest has been generated in association study or linkage disequilibrium (LD) mapping of complex diseases. The success of LD mapping depends heavily on the levels of LD between markers and genetic traits.<sup>1</sup> Positive selection can lead to an increase in the frequency of an advantageous allele and to high levels of LD in the vicinity of the trait gene.<sup>2–6</sup> Many forces determine disequilibrium on a genomic scale. Migration, nonrandom mating, variation in mutation rates, nonuniform recombination rates, and genetic drift, all immediately come to mind. Positive selection is one major force that increases LD locally rather than globally across the genome. A favorable mutation increases the frequency of the chromosome segment on which it occurs, and until that segment shrinks by recombination, neutral alleles at nearby loci will hitchhike to success. If the favorable mutation reaches fixation, then a selective sweep is declared.<sup>7–9</sup> Therefore, selection may shed light on complex diseases and human evolution. There is considerable interest in developing statistical methods to detect recent positive selection.<sup>10–20</sup>

Statistical geneticists studying positive selection have focused their efforts on two strategies: targeted examination and testing of candidate genes,<sup>21,22</sup> and data mining of genome association scans.<sup>23–25</sup> Methods to detect recent positive selection are either genotype-based or haplotype-based. The latter methods merit a brief summary.<sup>21,23,26,27</sup> Hanchard *et al.*<sup>26</sup> suggested screening for positive selection by passing a sliding window across a genomic region. At each position of the window, a test of haplotype similarity was evaluated. The extended haplotype homozygosity test proposed by Sabeti *et al.*<sup>21</sup> also assessed the age of a core haplotype. In view of the

computational burdens encountered in genome-wide scans, there has recently been a swing back to genotype-based methods. Tang *et al.*<sup>28</sup> proposed a homozygote counting method for genome scan data. This approach captured the decay of genotype homozygosity around a central single-nucleotide polymorphism (SNP). In contrast, the Sabeti *et al.*<sup>21</sup> statistic was designed to highlight the decay of extended haplotype homozygosity among extended haplotypes of a core haplotype. The counting method of Tang *et al.*<sup>28</sup> encouraged comparison of the homozygosity profiles of different populations. Although it is computationally fast, the counting method suffers from information loss, particularly with phase-known data.

Positive selection might cause LD and extended stretches of homozygosity.<sup>29</sup> These hitchhiking effects are most pronounced in genomic regions of low recombination. Although extended stretches of homozygosity also occur in inbreeding, the stretches occur randomly rather than systematically across the genome. In fact, none of the other forces that disrupt genetic equilibrium function in the targeted way of positive selection. For this reason, geneticists have been anxious to study the homozygous tracts of the human genome. Gibson *et al.*<sup>30</sup> examined the length, number, and distribution of homozygous tracts of SNPs among the HapMap reference populations without much theoretical analysis. In this article, we developed homozygosity score statistics to detect positive selection. These statistics were similar to the Tang *et al.*<sup>28</sup> statistic in that they rely on the length of homozygosity around each core SNP. We went beyond the analysis of Tang *et al.*<sup>28</sup> and calculated the mean and variance of each statistic under the appropriate null hypothesis. This facilitated computation of  $P$ -values by a normal approximation.

<sup>1</sup>Department of Statistics, The Texas A&M University, College Station, TX, USA; <sup>2</sup>Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA; <sup>3</sup>Department of Epidemiology, MD Anderson Cancer Center, University of Texas, Houston, TX, USA

\*Correspondence: Dr R Fan, Department of Statistics, The Texas A&M University, 447 Blocker Building, College Station, TX 77843-3143, USA. Tel: +1 979 845 3152 or 3141 (main office); Fax: +1 979 845 3144; E-mail: rfan@stat.tamu.edu

Received 23 June 2009; revised 4 January 2010; accepted 26 March 2010; published online 12 May 2010

Our three tests included (a) an extended genotype-based homozygosity score test (EGHST), (b) a hidden Markov model score test (HMMST), and (c) an extended haplotype-based homozygosity score test (EHHST). The null hypothesis of EGHST unrealistically postulated both Hardy–Weinberg equilibrium (HWE) and linkage equilibrium. The EHHST explicitly took into account multilocus LD. The HMMST occupied the intermediate ground of allowing for pairwise LD. In short, the EHHST was the most conservative test. We derived the tests and investigated their type I errors by simulation. We then focused on EHHST as it is the most robust. Under several demographic population models, we evaluated, by simulation, the fact that EHHST leads to appropriate false-positive rates. We investigated the power of EHHST by comparing it with popular methods.<sup>3,11,12,17,21,26</sup> It was found that EHHST has a higher or similar power as the existing popular methods. We also applied the tests to the previously studied HapMap Phase II data. Our results were consistent with previous findings across the genome and within specific candidate regions. We identified new candidate regions that were not reported before and were close to those reported previously.

## MATERIALS AND METHODS

Consider a random sample of  $n$  unrelated individuals typed on a large number of SNPs. Assume that the core SNP is SNP 0, which is the central SNP. Hence, the SNPs around the core SNP 0 can be denoted as  $k = \dots, -2, -1, 0, 1, 2, \dots$ . Here, one may need to truncate if the core SNP 0 is on the boundary or is close to the boundary. Let  $M$  be the indicator of whether SNP 0 is homozygous, let  $L$  be the number of consecutive homozygous SNPs flanking SNP 0 on the left, and let  $R$  be the number of consecutive homozygous SNPs flanking SNP 0 on the right. If the core SNP 0 is heterozygous ( $M=0$ ), then we define  $L=R=0$ . The extent of homozygosity is measured by the total,  $T=L+M+R$ . The quantities  $L$ ,  $M$ ,  $R$ , and  $T$  are random variables that vary from person to person. If we can find the mean  $\mu$  and variance  $\sigma^2$  of  $T$ , then we can conduct a test for excess homozygosity. More precisely, let  $T_j$  be the value of  $T$  for person  $j$  in the random sample. On the basis of the central limit theorem, the score statistic

$$S = \frac{1}{\sqrt{n\sigma^2}} \sum_{j=1}^n (T_j - \mu) \quad (1)$$

should be approximately standard normal. In this article, we consider three tests: EGHST, HMMST, and EHHST. As we are concerned with excess homozygosity, a one-sided test is appropriate.

Before we calculate the mean and variance of each test statistic, let us consider their corresponding null hypotheses. In each instance, the null hypothesis includes random mating and hence global HWE. Thus, the frequency of phased haplotypes  $H_1/H_2$  is  $2h_1h_2$  when  $H_1 \neq H_2$ , and  $h_1^2$  when  $H_1=H_2$ . Here,  $h_1$  and  $h_2$  are frequencies of haplotypes  $H_1$  and  $H_2$ , respectively. Only the null hypothesis of EGHST invokes the further assumption of linkage equilibrium; here,  $h_1$  and  $h_2$  equal the product of the underlying allele frequencies. Under the null hypothesis of HMMST, SNPs exhibit pairwise but not higher-order LD. For EHHST, arbitrary LD is allowed. In summary, the null hypotheses of the three tests are the following:

- Null hypothesis of EGHST: HWE and linkage equilibrium;
- Null hypothesis of HMMST: HWE and pairwise LD, but no higher-order disequilibrium interactions;
- Null hypothesis of EHHST: HWE and arbitrary multilocus LD.

In human genome, LD tends to extend the stretch of homozygosity surrounding a central marker, given high-density SNPs such as the HapMap Phase II data. The mean  $\mu$  calculated for the EGHST is consequently too small, and the EGHST is anticonservative. In other words, there are too many false positives favoring selection. As the other extreme, EHHST tends to condition on existing haplotype diversity and is very conservative. The HMMST stands between these extremes and conditions on pairwise LD. Given the ubiquity of

pairwise disequilibrium, this seems to be a reasonable compromise. Regardless of the test, one can decompose the theoretical mean of  $T$  as  $\mu = E(L) + E(M) + E(R)$ . Because  $M$  is an indicator random variable,  $E(M) = \Pr(M=1)$  and  $\text{Var}(M) = \Pr(M)[1 - \Pr(M)]$ . If we let  $X_k$  to be the unordered genotype of SNP  $k$ , then it is natural to calculate  $E(L)$  as  $E[E(L|X_0)]$ . As  $X_0$  takes only three possible values, the outer expectation in  $E[E(L|X_0)]$  is trivial to compute. The case  $X_0=1/2$  is easiest of all, because  $L=0$  when  $X_0=1/2$  and  $M=0$ . Similar comments apply to  $E(R)$ . The most natural route to calculate variance  $\sigma^2$  follows the formula

$$\begin{aligned} \text{Var}(T) &= \text{Var}(L) + \text{Var}(M) + \text{Var}(R) \\ &\quad + 2 \text{Cov}(L, M) + 2 \text{Cov}(L, R) + 2 \text{Cov}(M, R). \end{aligned}$$

Again, it is productive to condition on  $X_0$ . For instance,

$$\begin{aligned} \text{Var}(L) &= \text{Var}[E(L|X_0)] + E[\text{Var}(L|X_0)], \\ \text{Var}(R) &= \text{Var}[E(R|X_0)] + E[\text{Var}(R|X_0)], \end{aligned}$$

and, assuming  $L$  and  $R$  are independent given  $X_0$ ,

$$\begin{aligned} \text{Cov}(L, R) &= \text{Cov}[E(L|X_0), E(R|X_0)] + E[\text{Cov}(L, R|X_0)] \\ &= \text{Cov}[E(L|X_0), E(R|X_0)]. \end{aligned}$$

It is also worth pointing out that  $E(LM) = E(L)$  and  $E(RM) = E(R)$ , because  $L$  and  $R$  equal 0 when  $M$  does, and when  $M=1$ ,  $LM$  equals  $L$  and  $RM$  equals  $R$ . Thus, one has

$$\begin{aligned} \text{Cov}(L, M) &= E(LM) - E(L)E(M) = E(L)[1 - E(M)], \\ \text{Cov}(R, M) &= E(RM) - E(R)E(M) = E(R)[1 - E(M)]. \end{aligned}$$

These considerations emphasize the importance of finding the distributions of  $L$  and  $R$  conditional on  $X_0=1/1$  and  $X_0=2/2$ . The next few sections tackle this issue.

### The distribution of $L$ and $R$ under the null hypothesis of EGHST

Under the dual assumptions of HWE and linkage equilibrium, the conditional distributions of the random variables  $L$  and  $R$  depend only on  $M$  and not on the particular value of  $X_0$ . Let  $p_{k1}$  and  $p_{k2}$  be the frequencies of the two alleles at SNP  $k$ . In this notation, one can readily deduce that

$$\begin{aligned} \Pr(M = 1) &= p_{01}^2 + p_{02}^2, \\ \Pr(R \geq r | M = 1) &= \prod_{k=1}^r (p_{k1}^2 + p_{k2}^2), \\ \Pr(L \geq \ell | M = 1) &= \prod_{k=-\ell}^{-1} (p_{k1}^2 + p_{k2}^2), \end{aligned}$$

where the products are empty when  $r=0$  or  $\ell=0$ . In practice, one can either estimate the allele frequencies  $p_{k1}$  and  $p_{k2}$  from the sample or substitute known values for them. To compute the conditional means and variances of  $L$  and  $R$  numerically, we recommend the right-tail sums

$$E(Y) = \sum_{j=1}^{\infty} \Pr(Y \geq j), \quad E(Y^2) = \sum_{j=1}^{\infty} (2j - 1) \Pr(Y \geq j), \quad (2)$$

valid for any nonnegative random variable  $Y$  with integer values. The sums defining  $E(Y)$  and  $E(Y^2)$  can be truncated as soon as they stabilize.

### The distribution of $L$ and $R$ under the null hypothesis of HMMST

To find the conditional distributions of  $L$  and  $R$  under this scenario, we run a Markov chain, the states of which are the three unordered SNP genotypes 1/1, 1/2, and 2/2 and the epochs of which are SNPs. If we again hypothesize that SNP 0 is the central SNP, then the genotype sequence  $\dots, X_{-1}, X_0, X_1, \dots$  constitutes the chain. Every SNP emits a signal, either a 1 for a homozygote or a 0 for a heterozygote. Assuming pairwise LD but no higher-order linkage interactions, the two sections of the chain to the left and right of the central SNP are independent, conditional on the state  $X_0$  at that SNP. The only nontrivial states  $X_0$  that come into effect at SNP 0 are 1/1 and 2/2, and these occur with the Hardy–Weinberg probabilities  $p_{01}^2$  and  $p_{02}^2$ .

To compute the conditional mean and variance of  $R$ , it suffices to compute the probabilities  $\Pr(R \geq r | X_0)$ . This can be achieved by running Baum's forward algorithm for an infinite sequence of emitted 1s. One pass of the algorithm is adequate. When SNP  $r$  is visited,  $\Pr(R \geq r | X_0)$  becomes available. This description omits the mention of transition probabilities. Along either haplotype, the transition from allele  $j$  at SNP  $r$  to allele  $k$  at SNP  $r+1$  is governed by the known conditional probabilities that explicitly account for pairwise LD. These conditional probabilities can be readily estimated from sample data. We traverse the left and right sections in opposite direction. Hence, their transition probabilities must take this into account.

Under the assumption of no genotyping error, the complexities of the hidden Markov chain can be replaced by simple recurrence relations. Let  $p_{r,j \rightarrow k}$  be the LD probability that allele  $j$  at locus  $r$  is followed by allele  $k$  at locus  $r+1$  on a chromosome segment containing both loci. If we also let

$$a_{rj} = \Pr(X_r = j | j, R \geq r | X_0 = 1/1), \tag{3}$$

then we can conclude that  $\Pr(R \geq r | X_0 = 1/1) = a_{r1} + a_{r2}$ . Thus, computing the conditional mean and variance of  $R$  reduces the problem to computation of  $a_{r1}$  and  $a_{r2}$ . By convention, we consider  $a_{01} = 1$  and  $a_{02} = 0$ . These choices lead to the recurrences

$$\begin{aligned} a_{r1} &= \Pr(X_r = 1/1, R \geq r | X_0 = 1/1) \\ &= \Pr(X_r = 1/1, R \geq r-1, X_{r-1} = 1/1 \text{ or } 2/2 | X_0 = 1/1) \\ &= \Pr(R \geq r-1, X_{r-1} = 1/1 | X_0 = 1/1) \Pr(X_r = 1/1 | X_{r-1} = 1/1) \\ &\quad + \Pr(R \geq r-1, X_{r-1} = 2/2 | X_0 = 1/1) \Pr(X_r = 1/1 | X_{r-1} = 2/2) \\ &= a_{r-1,1} (p_{r-1,1 \rightarrow 1})^2 + a_{r-1,2} (p_{r-1,2 \rightarrow 1})^2, \\ a_{r2} &= a_{r-1,1} (p_{r-1,1 \rightarrow 2})^2 + a_{r-1,2} (p_{r-1,2 \rightarrow 2})^2. \end{aligned}$$

Computation of the vector  $a_r = (a_{r1}, a_{r2})$  should continue until

$$(2r-1) \Pr(R \geq r | X_0 = 1/1) < \varepsilon$$

for  $\varepsilon > 0$  suitably small. To compute  $\Pr(R \geq r | X_0 = 2/2)$ , we similarly define

$$b_{rj} = \Pr(R \geq r, X_r = j | j | X_0 = 2/2). \tag{4}$$

The  $b_{rj}$  satisfies exactly the same recurrences as  $a_{rj}$ , but differ in the initial conditions  $b_{01} = 0$  and  $b_{02} = 1$ .

As just stated, the distribution of  $L$  is independent of  $R$ , given  $X_0$ . Let  $c_{lj}$  be the probability that  $X_{-l} = j | j$  and  $L \geq l$ , given  $X_0 = 1/1$ . The conventions  $c_{01} = 1$  and  $c_{02} = 0$  are consistent with the formula  $\Pr(L \geq l | X_0 = 1/1) = c_{l1} + c_{l2}$ . Furthermore, we have recurrences

$$\begin{aligned} c_{l1} &= \Pr(X_{-l} = 1/1, L \geq l | X_0 = 1/1) = \Pr(X_{-l} = 1/1, L \geq l-1, X_{-l+1} = 1/1 \text{ or } 2/2 | X_0 = 1/1) \\ &= \Pr(L \geq l-1, X_{-l+1} = 1/1 | X_0 = 1/1) \Pr(X_{-l} = 1/1 | L \geq l-1, X_{-l+1} = 1/1, X_0 = 1/1) \\ &\quad + \Pr(L \geq l-1, X_{-l+1} = 2/2 | X_0 = 1/1) \Pr(X_{-l} = 1/1 | L \geq l-1, X_{-l+1} = 2/2, X_0 = 1/1) \\ &= c_{l-1,1} \Pr(X_{-l} = 1/1 | X_{-l+1} = 1/1) + c_{l-1,2} \Pr(X_{-l} = 1/1 | X_{-l+1} = 2/2) \\ &= \frac{\Pr(X_{-l} = 1/1)}{\Pr(X_{-l+1} = 1/1)} c_{l-1,1} (p_{l-1,1 \rightarrow 1})^2 + \frac{\Pr(X_{-l} = 1/1)}{\Pr(X_{-l+1} = 2/2)} c_{l-1,2} (p_{l-1,2 \rightarrow 1})^2, \\ c_{l2} &= \frac{\Pr(X_{-l} = 2/2)}{\Pr(X_{-l+1} = 1/1)} c_{l-1,1} (p_{l-1,1 \rightarrow 2})^2 + \frac{\Pr(X_{-l} = 2/2)}{\Pr(X_{-l+1} = 2/2)} c_{l-1,2} (p_{l-1,2 \rightarrow 2})^2. \end{aligned}$$

If we let  $d_{lj}$  be the probability that  $L \geq l$  and  $X_{-l} = j | j$ , given  $X_0 = 2/2$ , then the same recurrences as  $c_{lj}$  hold for  $d_{lj}$  but the initial conditions are given by  $d_{01} = 0$  and  $d_{02} = 1$ .

**The distribution of  $L$  and  $R$  under the null hypothesis of EHHST**

In the presence of arbitrary LD, the fast recurrences (3) and (4) for  $a_{rj}$  and  $b_{rj}$  no longer apply. However, if we define  $h_{i_0, \dots, i_r}$  to be the population frequency of the haplotype  $(i_0, \dots, i_r)$  extending from SNP 0 to SNP  $r$ , then the formula

$$\Pr(R \geq r | X_0 = i_0 / i_0) = \frac{1}{P_{0,i_0}^2} \sum_{i_1=1}^2 \dots \sum_{i_r=1}^2 h_{i_0, \dots, i_r}^2$$

delivers the required right-tail probabilities. When all conceivable haplotypes are possible, there are  $2^r$  terms in the multiple sum, and the formula as it stands is cumbersome. On the other hand, if only a few haplotypes are possible, then the sum is straightforward to evaluate. Moment formulas (2) are still applicable. The haplotype frequencies  $h_{i_0, \dots, i_r}$  can be estimated from genotype data by the EM algorithm.<sup>31,32</sup>

**RESULTS**

**Type I error rates**

By construction, EHHST was the most conservative among the three tests of EGHST, HMMST, and EHHST. For further confirmation, we performed false-positive (type I) error comparison by simulating genotype data under the null of EGHST. The results were reported in Supplementary I, in which we showed that the type I error rates of EHHST were the smallest. Thus, EHHST was the most robust among the three. Hereafter, we focused our attention on evaluating the performance of EHHST.

We first used SelSim to simulate data under the neutral model.<sup>33</sup> In a genomic region, a few fixed numbers 51, 61, 71, 81, 91, and 101 of SNPs were simulated to evaluate type I error rates. In addition, uniform recombination rates of  $\rho = 1.5, 3, 6,$  and  $9$  between SNPs were assumed. To calculate an empirical type I error rate, 5000 random samples of  $n = 60$  or  $100$  individuals were generated. For each sample, an empirical EHHST value for the central SNP was calculated. The type I error rates at two nominal levels  $\alpha = 0.05$  and  $0.01$  were reported in Table 1, which were the proportion of the EHHST values that exceeded the 95th and 99th percentiles of the standard normal. When the number of SNPs is 51, the type I error was much bigger than the nominal levels for any of the four recombination rates. Interestingly, type I error rates decreased when the number of SNPs increased to 61 and then to 71, and the trend continued until the number of SNPs increased to 71. Once the number of SNPs reached 71, the type I error rates stabilized. Hence, truncation at the boundary SNPs caused a problem of high false positives. Fortunately, almost all contemporary genomic data comprised a large number of SNPs. On the basis of the results of Table 1, the type I error rates were lower than or around the nominal level, except for the recombination rate  $\rho = 1.5$  when the number of SNPs was larger or equal to 71. When  $\rho = 1.5$ , type I error rates were generally higher than the nominal levels but not very high. Therefore, EHHST had appropriate type I error rates when it was used to calculate the test score of SNPs that are reasonably far away from the boundary ( $\geq 35$ ).

**Table 1 Type I error rates of the extended haplotype-based homozygosity score test (EHHST)**

Sample Size	No. of n	No. of SNPs	Nominal level		No. of SNPs	Nominal level							
			$\rho$	$\alpha = 0.05$		$\alpha = 0.01$	$\rho$	$\alpha = 0.05$	$\alpha = 0.01$				
100	51	1.5	0.1496	0.0592	61	1.5	0.0832	0.0260					
			0.1268	0.0486			0.0548	0.0164					
			0.1066	0.0388			0.0462	0.0130					
		9	0.1036	0.0362		0.0464	0.0124						
			71	1.5		0.0592	0.0182	81	1.5	0.0562	0.0148		
						0.0392	0.0092			0.0374	0.0080		
	0.0362	0.0062			0.0272	0.0052							
	91	1.5	0.0296	0.0042	0.0042	9	0.0306	0.0048					
				101	1.5			0.0566	0.0184	101	1.5	0.0578	0.0200
												0.0414	0.0102
		0.0260	0.0052			0.0326	0.0056						
		60	91	1.5	0.0262	0.0054	9	0.0274	0.0052				
3					0.0580	0.0208			101	1.5	0.0626	0.0226	
	0.0344										0.0084	0.0396	0.0112
	0.0298		0.0052	0.0330			0.0084						
9	0.0252		0.0062	9	0.0248	0.0052							

All results were based on 5000 simulations using Software SelSim.<sup>33</sup>

To investigate the impact of demographic population history on EHHST, we performed coalescent simulations using *ms*.<sup>34</sup> We evaluated the type I error rates of EHHST under a few plausible population genetic demographic models. Specifically, we considered four demographic models that are similar to those considered in Hanchard *et al*.<sup>26</sup>

1. *Population structure*: Two equal-sized sub-populations were simulated, which exchanged migrants with a probability 0.1;
2. *Population expansion*: A rapid population growth was simulated with a current population size of 10 000, and the population had a constant population size until 500 generations ago when it expanded exponentially by a factor of 100 to reach the current day population size of 10 000;
3. *Population bottleneck 150/300*: A panmictic population was simulated, which had a constant size of 10 000 until  $T_1=300$  generations ago, when it underwent an instantaneous size reduction to 5000, followed by a period of 150 generations of constant size, followed by a rapid exponential population expansion in the last  $T_2=150$  generations to reach a current day size of 20 000;
4. *Population bottleneck 250/500*: A population similar to the above *Population bottleneck 150/300*: Except that  $T_1=500$  and  $T_2=250$ .

Again, a genomic region of 101 SNPs was simulated with four recombination fractions  $\rho=1.5, 3, 6,$  and  $9$ . In addition, 5000 samples of  $n=60$  or  $100$  were generated to calculate the empirical type I error rates one by one. The results were reported in Table 2. When the recombination fractions are  $3, 6,$  or  $9$ , type I error rates were lower or around the nominal levels. Similar to the results of Table 1, the type I error rates were generally higher than the nominal levels when  $\rho=1.5$ . For the four models, the type I error rates of our EHHST were lower than those of Hanchard's HS reported in Hanchard *et al*,<sup>26</sup> p155, second paragraph of the left column. The EHHST was reasonably robust for the four simple demographic models.

### Power of EHHST

To perform power comparisons with the existing methods, we focused our attention on the results reported in Figure 1 of Hanchard *et al*.<sup>26</sup> The figure contained a comparison of Hanchard's HS, Sabeti's EHH, Tajima's *D*-test, Fu and Li's *D*-test, Fay and Wu's *H*-test, and Hudson's haplotype-partition method.<sup>3,11,12,17,21,26</sup> By a comprehensive and careful comparison, Hanchard *et al*<sup>26</sup> concluded that Hanchard's HS and Sabeti's EHH are the two best tests. Thus, we mainly compared the performance of our EHHST with Hanchard's HS and Sabeti's EHH.

Similar to the study by Hanchard *et al*,<sup>26</sup> we performed coalescent simulations for power comparison by SelSim.<sup>33</sup> First, all parameters were taken as exactly those of Figure 1 of Hanchard *et al*,<sup>26</sup> with one exception: we simulated a genomic region comprising 101 SNPs instead of 50, to avoid a potential problem caused by truncation at the boundary (refer to type I error rates). For readers' convenience, let us briefly describe the models and parameters as follows. In Hanchard *et al*,<sup>26</sup> three different uniform recombination rates,  $\rho=4N_0r=1.5, 3,$  and  $6$ , between SNPs were used in the simulation, and three different allele frequencies (0.1, 0.2, and 0.4) were used for the minor allele of the central SNP. Here,  $N_0$  is the diploid population size and  $r$  is the probability of crossover per generation between SNPs. In our simulations, we used four recombination rates  $\rho=1.5, 3, 6,$  and  $9$ , and six present day population frequencies of the derived allele for the central SNP (0.1, 0.2, 0.4, 0.6, 0.8, and 0.9). As in Hanchard *et al*,<sup>26</sup> a partial selective sweep was assumed for the central SNP by using a selection coefficient  $s=500$ .

To calculate an empirical power level, we simulated 5000 samples of 200 chromosomes or  $n=100$  individuals. For each sample, we calculated an empirical EHHST value for the central SNP. Thereafter, empirical power was calculated as the proportion of the 5000 EHHST values that exceeded the 95th and 99th percentiles of the standard normal. The results are reported in Table 3. At the nominal level  $\alpha=0.05$ , the empirical power of EHHST was higher than 0.9410, irrespective of the four recombination rates and the five present day

**Table 2** Type I error rates of the extended haplotype-based homozygosity score test (EHHST)

Demographic model	Sample size, n	$\rho$	Nominal level		Sample size, n	$\rho$	Nominal level	
			$\alpha=0.05$	$\alpha=0.01$			$\alpha=0.05$	$\alpha=0.01$
Population structure	100	1.5	0.0512	0.0154	60	1.5	0.0538	0.0186
		3	0.0406	0.0082		3	0.0380	0.0118
		6	0.0292	0.0060		6	0.0300	0.0064
		9	0.0264	0.0046		9	0.0262	0.0060
Population expansion	100	1.5	0.0664	0.0212	60	1.5	0.0514	0.0142
		3	0.0428	0.0094		3	0.0352	0.0090
		6	0.0414	0.0094		6	0.0272	0.0048
		9	0.0324	0.0058		9	0.0232	0.0042
Population bottleneck 150/300	100	1.5	0.1040	0.0384	60	1.5	0.0954	0.0346
		3	0.0516	0.0180		3	0.0516	0.0170
		6	0.0366	0.0082		6	0.0362	0.0092
		9	0.0310	0.0066		9	0.0302	0.0068
Population bottleneck 250/500	100	1.5	0.0864	0.0292	60	1.5	0.0870	0.0324
		3	0.0492	0.0140		3	0.0496	0.0154
		6	0.0348	0.0082		6	0.0324	0.0102
		9	0.0276	0.0072		9	0.0278	0.0084

All results were based on 5000 simulations using software *ms*,<sup>3</sup> and a genomic region of 101 SNPs was simulated.

**Table 3 Power of the extended haplotype-based homozygosity score test (EHHST)**

Selection coefficient	Sample size, n	Recombination rates, $\rho$	Nominal level, $\alpha$	Present day population frequency of derived allele						
				0.1	0.2	0.4	0.6	0.8	0.9	
s=500	100	1.5 <sup>#</sup>	0.05	0.9986	1.0000	1.0000	1.0000	1.0000	0.9998	0.9252
			0.01	0.9976	1.0000	1.0000	1.0000	1.0000	0.9968	0.7346
		3 <sup>#</sup>	0.05	0.9950	1.0000	1.0000	1.0000	1.0000	0.9988	0.7612
			0.01	0.9762	0.9998	1.0000	1.0000	1.0000	0.9820	0.4152
		6 <sup>#</sup>	0.05	0.9718	1.0000	1.0000	1.0000	1.0000	0.9984	0.7750
			0.01	0.8898	0.9998	1.0000	1.0000	1.0000	0.9806	0.4586
	9	0.05	0.9410	1.0000	1.0000	1.0000	1.0000	0.9976	0.7694	
		0.01	0.8026	0.9986	1.0000	1.0000	1.0000	0.9830	0.4874	
	60	1.5	0.05	0.9644	0.9994	1.0000	1.0000	0.9656	0.6276	
			0.01	0.9124	0.9970	1.0000	0.9998	0.8326	0.3236	
		3	0.05	0.8572	0.9988	1.0000	1.0000	0.9042	0.4328	
			0.01	0.7294	0.9902	1.0000	0.9980	0.7030	0.1790	
		6	0.05	0.7912	0.9912	1.0000	1.0000	0.9308	0.4640	
			0.01	0.5990	0.9638	1.0000	0.9988	0.7782	0.2198	
	9	0.05	0.6938	0.9798	1.0000	1.0000	0.9234	0.4618		
		0.01	0.4504	0.9244	0.9996	0.9992	0.7666	0.2104		

All results were based on 5000 simulations using Software SelSim.<sup>33</sup> The rows marked by # contain results that were calculated using the same models and parameters as those of Figure 1 of Hanchard *et al.*<sup>26</sup>

population frequencies 0.1, 0.2, 0.4, 0.6, and 0.8 of the derived allele of the central SNP. Most of the EHHST empirical power levels were around 0.98 at the nominal level  $\alpha=0.05$ . For the present day population frequency 0.9 of the derived allele of the central SNP, the empirical power of the EHHST was higher than 0.76. To compare the performance of our EHHST with Sabeti's EHH and Hanchard's HS, we showed the results of the power comparison in Figure 1. The two plots on the top of Figure 1, ie, EHH and HS plots, were taken from Hanchard *et al.*<sup>26</sup> Figure 1. The results of Figure 1 and Table 3 clearly show that EHHST performed just as well as or even better than Hanchard's HS and Sabeti's EHH.<sup>26</sup> One may want to notice that the power levels of Hanchard's HS and Sabeti's EHH reported in Figure 1 of Hanchard *et al.*<sup>26</sup> can be lower than or around 0.80 when allele frequency was 0.1 at the nominal level  $\alpha=0.05$ , although the rest of the power levels were larger than 0.90. The empirical power of EHHST, on the other hand, was high at the nominal level  $\alpha=0.05$ , with a minimum 0.9718 when  $\rho=6$ , and population frequency of the derived allele was equal to 0.1 for the five present day population frequencies 0.1, 0.2, 0.4, 0.6, and 0.8 of the derived allele of the central SNP, and  $\rho=1.5, 3$ , and 6. The empirical power of EHHST was high at the nominal level  $\alpha=0.01$ , with a minimum 0.8898 for the five derived allele frequencies 0.1, 0.2, 0.4, 0.6, and 0.8 of the central SNP, and  $\rho \leq 6$ .

In addition, we calculated empirical power by simulating 5000 samples of 120 chromosomes or  $n=60$  individuals. The HapMap data contained samples of size 60, and our results provided some insight into the samples. The results are reported in Table 3. The EHHST provided reasonably high power in this case for the five present day population frequencies 0.1, 0.2, 0.4, 0.6, and 0.8 of the derived allele of the central SNP. For the present day population frequency 0.9 of the derived allele, the empirical power of EHHST could be low.

To perform power comparison for the three tests proposed, we simulated data under the same conditions as those of Table 3 by SelSim to calculate the empirical powers of EGHST and HMMST. The results were reported in Tables 4 and 5 in Supplementary I. Again, 5000 samples of 200 chromosomes (ie,  $n=100$  individuals) or 120

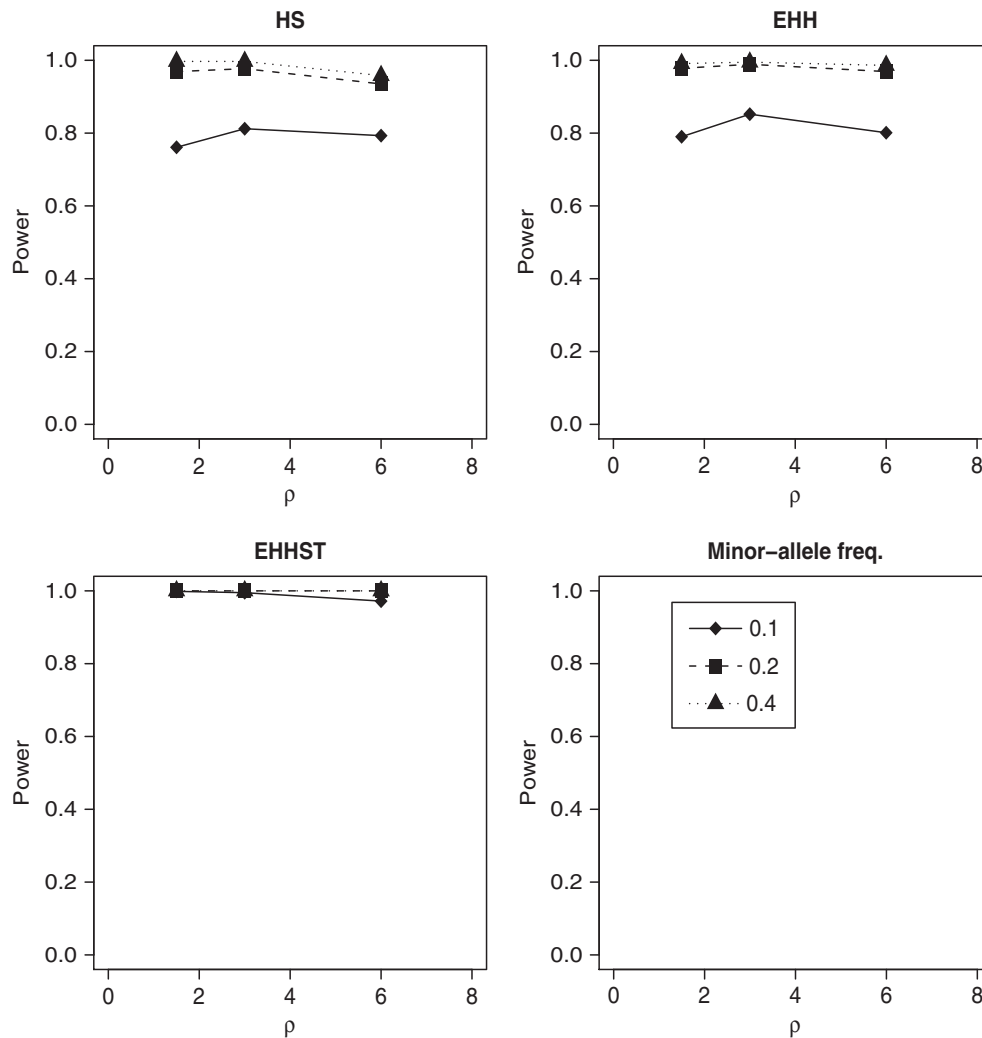
chromosomes (ie,  $n=60$  individuals) were simulated. As expected, the power of EGHST was higher than that of HMMST, which was generally more powerful than EHHST. Hence, EHHST was the most conservative.

#### HapMap Phase II data

We applied the proposed score test statistics to the whole-genome SNP data of HapMap Phase II.<sup>35</sup> Data sets include 3.1 million SNP genotypes from population samples of three continents: 60 CEPH Utah residents with ancestry from Northern and Western Europe (CEU); 60 Yorubas from Ibadan (YRI), Nigeria in Africa; and 45 Han Chinese from Beijing (CHB) and 45 Japanese from Tokyo (JPT), Japan, of Asia. The samples were downloaded from [http://www.hapmap.org/downloads/phasing/2007-08\\_rel22/phased/](http://www.hapmap.org/downloads/phasing/2007-08_rel22/phased/). The two Asian samples were combined into one, referred to hereafter as CHB+JPT, as instructed by the HapMap Consortium. We used only the unrelated individuals from the three samples, omitting the children in the trio families from the CEU and YRI samples.

#### Results in the candidate regions

To evaluate the performance of our proposed test statistics, we applied them to the HapMap Phase II data in the 20 autosome candidate regions that show strong signals in Table 1 of Sabeti *et al.*<sup>36</sup> Note that our tests were designed for autosome data (and we actually could not download sex-linked X and Y chromosome data in the above-mentioned HapMap website). Before we discuss our results in detail, let us give a rough summary: In 17 out of 20 candidates, EHHST values showed peaks for the selected population samples in column 2, Table 1 of Sabeti *et al.*<sup>36</sup> hence, there were extended stretches of homozygosity in these 17 regions, as a result of which positive selection could lead to excess homozygosity in the human genome. The three exceptions can be found in Figure 3 of Supplementary II: (a) a region around 78.3 Mb on chromosome 12, (b) the *BCAS3* gene region on chromosome 17, (c) the gene region of *CHST5*, *ADAT1*, and *KARS* on chromosome 16. In the following paragraphs, we limited our discussion to the candidate regions on chromosomes 2 and 15.



**Figure 1** Power comparison of EHHST with Sabeti's EHH and Hanchard's HS. The two plots on the top, ie, EHH and HS plots, were taken from Hanchard *et al*,<sup>26</sup> Figure 1.

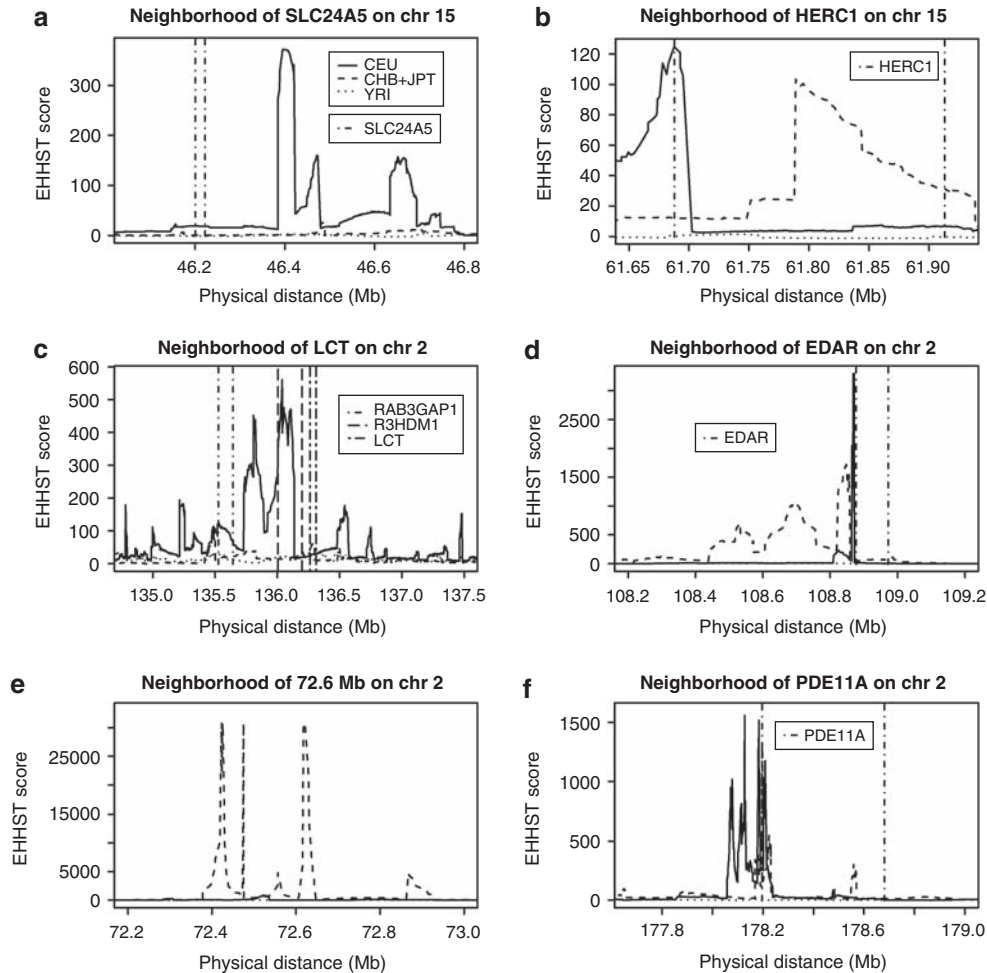
In particular, we considered the regions containing lactose tolerance gene *LCT* on chromosome 2 and the pigmentation gene *SLC24A5* on chromosome 15. For the sake of brevity, we treated the remaining candidate regions in Supplementary II.

Figure 2a explains an interesting fact about the *SLC24A5* gene on chromosome 15. The gene occurred between the two vertical dotted-dashed lines from 46.20 to 46.22 Mb in Figure 2a. The highest peak of EHHST occurred around 46.4 Mb, which was reported in Table 1 of Sabeti *et al*,<sup>36</sup> the EHHST values of CHB+JPT and YRI samples were very low and the test scores of the YRI sample were uniformly the lowest. Our results were consistent with those of Sabeti *et al*<sup>36</sup> and Lamason *et al*,<sup>37</sup> who argued for positive selection on the basis of a striking reduction in heterozygosity in the CEU sample. In a 200 kb region around gene *HERC1* on chromosome 15, the CHB+JPT sample showed signs of positive selection (Table 1, Sabeti *et al*<sup>36</sup>). The EHHST values were plotted in Figure 2b. Again, the gene was located between the vertical dotted-dashed lines, from 61.69 to 61.91 Mb on the Figure 2b. The EHHST values of CHB+JPT were clearly highest within most parts of the *HERC1* gene. Hence, the CHB+JPT sample showed long extended haplotype homozygosity in the gene region.

The *LCT* gene was sited between 136.26 and 136.32 Mb on chromosome 2, and LD extended about 3.2 Mb around it in the

CEU sample.<sup>38–40</sup> Two other genes were located in the same region, *RAB3GAP1* between 135.53 and 135.64 Mb and *R3HDM1* between 136.01 and 136.20 Mb. Our EHHST values plotted in Figure 2c were noticeably higher in the CEU sample than in YRI and CHB+JPT samples, confirming the previous results. Most striking of all was that the EHHST statistic spiked directly over gene *R3HDM1* right next to gene *LCT*. Although this did not prove positive selection, the fact that a mutation deregulating the *LCT* gene occurred on the conserved haplotype strongly favors this interpretation. Because of the high-density SNPs of HapMap data, high-degree LD may not necessarily be the selection signal. Long extended haplotype homozygosity, however, could lead to high EHHST values and interesting signals for further investigations.

Two other regions on chromosome 2, a 1.0 Mb region around gene *EDAR* and an 800 kb region around 72.6 Mb, showed strong evidence of selection in the CHB+JPT sample (Table 1, Sabeti *et al*<sup>36</sup>). The EHHST values plotted in Figure 2d–e confirmed the previous findings. The sharp EHHST peak for the CHB+JPT sample located very close to the *EDAR* region between 108.88 and 108.97 Mb (Figure 2d). In short, the EHHST statistic provided evidence of selection signal of the CHB+JPT sample in the region of the *EDAR* gene. In comparison, our EHHST values plotted in Figure 2e confirmed that the CHB+JPT



**Figure 2** The EHHST values of three population samples taken from HapMap Phase II data: Graph (a) in the region of the *SLC24A5* gene and Graph (b) in the region of the *HERC1* gene on chromosome 15, Graph (c) in the region of *RAB3GAP1*, *R3HDM1*, *LCT* genes, Graph (d) in the region of the *EDAR* gene, Graph (e) in the region around 72.6 Mb, and Graph (f) in the region of the *PDE11A* gene on chromosome 2. The vertical dotted-dashed legend indicates the locations of genes *SLC24A5*, *HERC1*, *EDAR*, and *PDE11A* in Graphs (a), (b), (d), and (f), respectively; the three vertical legends in Graph (c) indicate the locations of genes *RAB3GAP1*, *R3HDM1*, and *LCT*, which were sited at intervals of (135.53, 135.64), (136.01, 136.20), and (136.26, 136.31), respectively. chr, chromosome.

sample has a strong selection signal in the 800 kb region around 72.6 Mb. Interestingly, the EHHST values reached the highest in the region (Figure 2e and Table 2 of Supplementary III). In a 1.2-Mb region around gene *PDE11A*, both CHB+JPT and CEU samples were reported to have a strong selection signal (Table 1, Sabeti *et al*<sup>36</sup>). The EHHST peaks of CHB+JPT and CEU samples overlapped the *PDE11A* region in Figure 2f.

#### New candidate regions for further investigation based on the high EHHST values

Among the three proposed test statistics, EHHST was the most conservative. High EHHST values in a region indicated that there were long stretches of homozygosity. In the 20 candidate regions reported previously, we found that EHHST values show peaks in 17 of them. All these features encouraged us to use EHHST in search of new candidate regions for further investigations. Before selecting a candidate region, we first selected SNPs for natural selection as follows: (1) the selected SNP had a high EHHST value in the top one percentile, ie, the EHHST value of the SNP is in the top one percentile

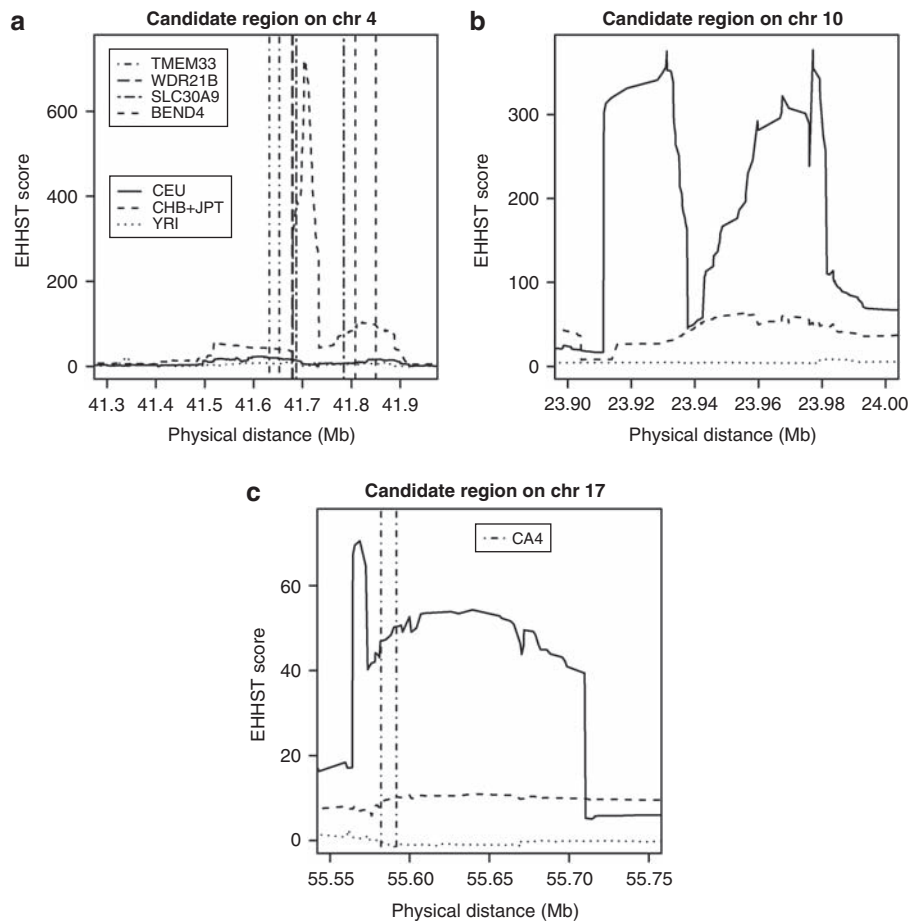
of all SNPs of a chromosome in which the SNP is located; (2) the selected SNP had an allele that is likely to be newly derived by using the data from <http://www.hg-wen.uchicago.edu/selection/frontpage.html> of the University of Chicago;<sup>23</sup> (3) the derived allele of the selected SNP had a high frequency that was larger than 0.5 in the tested population; (4) the derived allele of the selected SNP was likely to be highly differentiated among the three populations of CHB+JPT, CEU, and YRI, ie, the  $F_{st}$  score of the SNP was in the top one percentile of all  $F_{st}$  scores of SNPs on a chromosome.<sup>41–43</sup> A candidate region was selected if there was a long list of SNPs that satisfied the four selection criteria.

On the basis of the four criteria described above, 21 candidate regions were found for natural selection (Supplementary III). In the 21 candidate regions, 3 were close to regions reported in Sabeti *et al*,<sup>36</sup> and 12 were not reported; we counted these 15 regions as new candidates. The remaining six regions were within regions reported in Sabeti *et al*.<sup>36</sup> A brief description of the 15 new candidates is presented in Table 4. For the three regions that were close to regions reported in Sabeti *et al*,<sup>36</sup> we plotted the EHHST values in Figure 3.

**Table 4** New candidate regions for natural selection identified by the four criteria described in the main text and Supplementary II

Region	Chr	Tested population	Starting position (bp)	Ending position (bp)	Size (bp)	Number of SNPs	Genes in or near the region
1	1	CHB+JPT	75 329 244	75 512 920	183 676	11	<i>LHX8, SLC44A5</i>
2	2	CHB+JPT	17 200 074	17 265 872	65 798	19	
3	2	CHB+JPT	103 484 100	103 606 852	122 752	30	
4	3	CHB+JPT	106 178 646	106 306 013	127 367	22	
5 <sup>#</sup>	4	CHB+JPT	41 521 093	41 849 931	328 838	69	<i>SLC30A9, TMEM33, BEND4, WDR21B</i>
6	5	CHB+JPT	117 006 587	117 620 240	613 653	32	
7	7	CEU	119 168 428	119 230 892	62 464	9	
8	8	CEU	50 635 491	50 943 725	308 234	13	
9	8	CEU	52 876 153	52 926 708	50 555	16	<i>PCMTD1</i>
10 <sup>#</sup>	10	CEU	23 929 695	23 945 921	16 226	7	
11	11	CHB+JPT	38 066 457	38 263 161	196 704	15	
12	11	CEU	38 164 195	38 448 731	284 536	16	
13	12	CHB+JPT	42 674 621	42 879 128	204 507	12	<i>TMEM117</i>
14	13	CHB+JPT	63 174 790	63 348 614	173 824	17	
15 <sup>#</sup>	17	CEU	55 588 298	55 698 601	110 303	10	<i>CA4</i>

<sup>#</sup> marks regions that are close to a candidate region reported in Table 1, Sabeti *et al.*<sup>36</sup> The seventh column, Number of SNPs, gives number of SNPs that satisfied the four criteria.



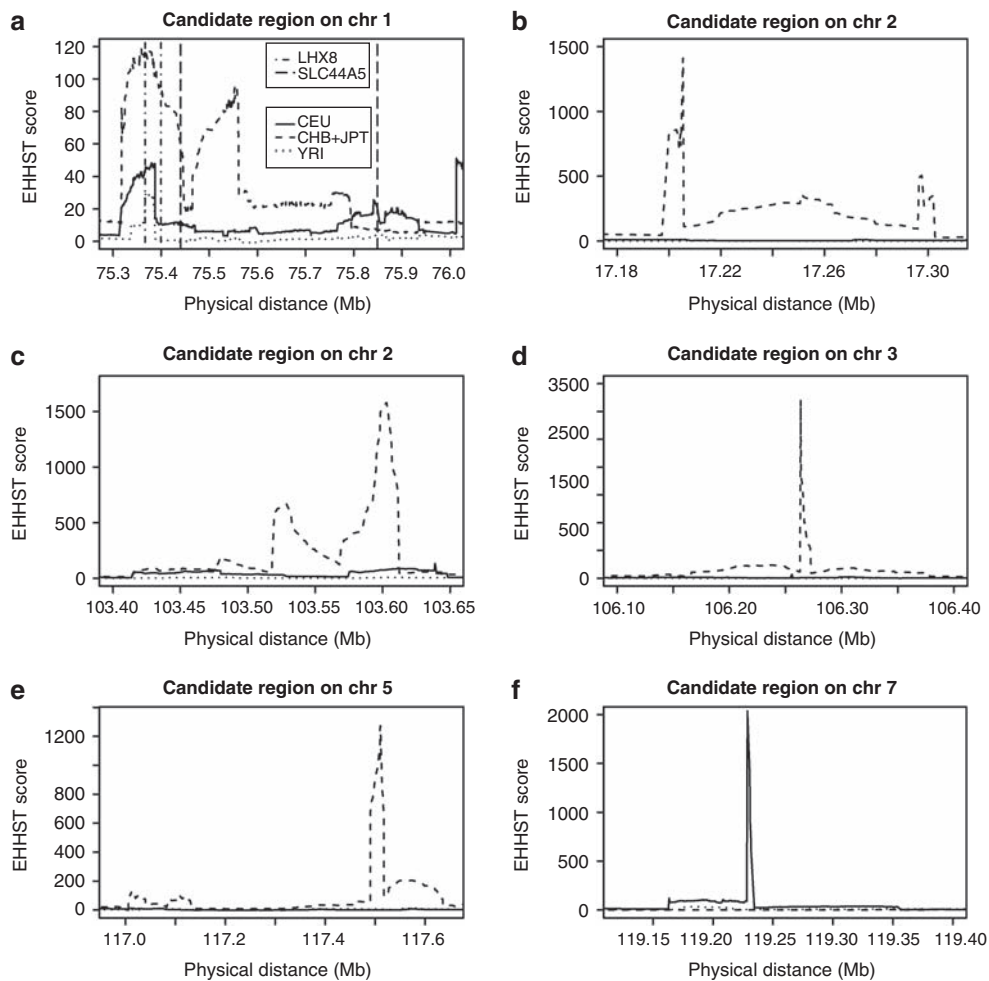
**Figure 3** The EHHST values of three population samples from HapMap Phase II data: Graph (a) in the region of *SLC30A9*, *TMEM33*, *BEND4*, and *WDR21B* genes on chromosome 4, Graph (b) in the candidate region on chromosome 10, and Graph (c) in the region of the *CA4* gene on chromosome 17. Gene locations are marked by vertical legends in Graph (a) for *TMEM33*, *WDR21B*, *SLC30A9*, and *BEND4*, which were sited at intervals of (41.63, 41.65), (41.678548, 41.679877), (41.69, 41.78), and (41.81, 41.85), respectively. The vertical dotted-dashed legend indicates the location of gene *CA4* in Graph (c).

The region containing the least number of SNPs satisfying the criteria (seven SNPs) was located at chromosome 10:23.9 Mb. It was reported because it was close to one candidate chr10:22.7 Mb of Sabeti *et al.*<sup>36</sup> Figure 3b shows the EHHST values of the three samples. It is clear that CEU sample has the highest EHHST values, which is consistent with the result in Table 4.

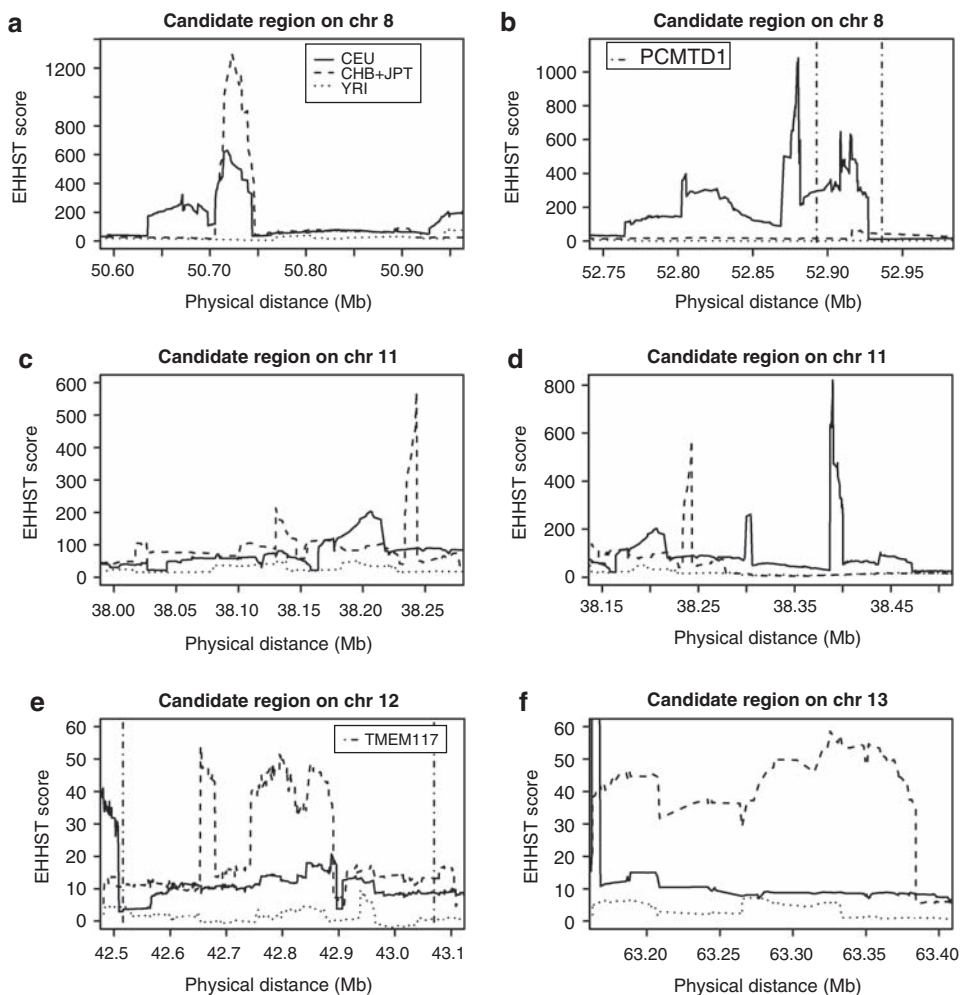
Other regions of Table 4 contain 9–69 SNPs that satisfy the four criteria. It was interesting that the region containing the most number of 69 SNPs was located on chromosome 4, from 41 521 093 to 41 849 931 bp, which overlapped with the candidate region chr4:42 Mb reported in Sabeti *et al.*<sup>36</sup> for natural selection in the CHB+JPT sample. Figure 3a shows that the EHHST values of the CHB+JPT sample were much higher than those of CEU and YRI samples in the gene region of *SLC30A9*, *TMEM33*, *BEND4*, and *WDR21B*, and the result is actually the same as that in Figure 1a of Supplementary II in the overlapped region. A region on chromosome 17 from 55 588 298 to 55 698 601 bp, which was close to the candidate region chr17:56.4 Mb in Sabeti *et al.*<sup>36</sup> was identified by our four criteria for natural selection in the CEU sample. Figure 3c shows that the EHHST values of the CEU sample were much higher than those of the other two samples. However, we failed to confirm the result of Sabeti *et al.*<sup>36</sup> in the chr17:56.4 Mb region of size 0.4 Mb by our three

tests (Figure 3b of Supplementary II). One possible reason for this discrepancy is that the distance/position of SNPs used in Sabeti *et al.*<sup>36</sup> was different from ours; if this is the case, then the number of new candidate regions that were not reported before is 12. In any case, it was encouraging to observe strong selection signals in 3+6=9 neighboring regions out of 21 by our methods, which were reported (or close to those reported) in Sabeti *et al.*<sup>36</sup>

It was interesting to study the 12 new regions in Table 4 for further dissection, which were not reported before and which were not close to the ones reported. In Figures 4 and 5, the EHHST values were plotted for a comparison in the 12 new regions. Figures 4a–e and 5e, f show that EHHST values of the CHB+JPT sample were either high or spiked over the regions on chromosomes 1, 2, 3, 5, 12, and 13 reported in Table 4. In particular, the EHHST values of the CHB+JPT sample spiked directly over the region of *LHX8* and *SLC44A5* genes in Figure 4a, and over the region of the *TMEM117* gene in Figure 5e. For the regions in Table 4 with natural selection signals in the CEU sample, the EHHST values of the CEU sample spiked over some parts of the regions on chromosomes 7 and 8 in Figure 4f and b, whereas the scores of CHB+JPT and YRI samples were very low. In addition, the EHHST values of the CEU sample spiked over the region of the *PCMTD1* gene in Figure 5b. The EHHST values of the CHB+JPT



**Figure 4** The EHHST values of three population samples from HapMap Phase II data: graph (a) in the region of *LHX8* and *SLC44A5* genes on chromosome 1, graphs (b, c) in the candidate regions on chromosome 2, graph (d) in the candidate region on chromosome 3, graph (e) in the candidate region on chromosome 5, and graph (f) in the candidate region on chromosome 7. In graph (a), the vertical dotted-dashed legend indicates the location of gene *LHX8*, and the vertical dashed legend indicates the location of gene *SLC44A5*.



**Figure 5** The EHHST values of three population samples from HapMap Phase II data: graph (a) in the candidate region and graph (b) in the region of the *PCMTD1* gene on chromosome 8, graphs (c, d) in the candidate regions on chromosome 11, graph (e) in the region of *TMEM117* gene on chromosome 12, and graph (f) in the candidate region on chromosome 13. The vertical dotted-dashed legend indicates the location of gene *PCMTD1* in graph (b) and the location of gene *TMEM117* in graph (e).

sample are generally the highest in Figure 5c, and the EHHST values of the CEU sample are generally the highest in Figure 5d, except in short parts of the regions. In short, the results of Figure 5c and d are consistent with those of chromosome 11 in Table 4. In Figure 5a, the EHHST values of both the CHB+JPT and CEU samples are high, whereas those of YRI are very low in the region between 50 635 491 and 50 943 725 bp, although CHB+JPT was not selected in Table 4 for the region.

#### Genome-wide scans of HapMap Phase II data

Given the encouraging results with the candidate regions, we performed a genome-wide scan of the HapMap Phase II data. The scan generated several results. In Supplementary IV, we present the results of chromosome 2 data. The features of the results of the remaining chromosomes were similar. The results of chromosome 2 are reported as three figures for the three tests – EHHST, HMMST, and EGHST. Each figure plots the test scores of one test for the CEU sample *versus* the YRI sample, one for the CHB+JPT sample *versus* the YRI sample, and one for the CEU sample *versus* the CHB+JPT sample. For the EGHST and HMMST statistics, scores were the highest for the

CHB+JPT sample and lowest for the YRI sample, with the CEU sample having intermediate scores (Figures 2 and 3 of Supplementary IV). This result was consistent with the finding by Gibson *et al*<sup>30</sup> that the YRI sample had the fewest long tracts of homozygosity. It was also consistent with current thinking about the demographic history of the three populations.

As a reflection of LD, EGHST and HMMST values were high across the genome (Figures 2 and 3 of Supplementary IV). Because the HMMST values adjust for pairwise LD, they were roughly half as high as EGHST values. By contrast, the EHHST values were generally low, with sharp spikes in just a few regions (Figure 1 of Supplementary IV). Hence, HWE was valid for most part of the genome; the high EGHST and HMMST values were most likely due to LD among SNP markers. From the plot of the high EHHST values, one could easily spot these narrow chromosome regions where HWE broke down.

#### Software and computational performance

Our C++ code for the proposed statistics is freely available on request to Dr Fan. The EGHST and HMMST are very fast computationally, taking only minutes to analyze a typical chromosome of the HapMap

Phase II data. In contrast, it requires hours per chromosome to compute the EHHST values. Hence, the EHHST seems to be the most suited for fine mapping in candidate gene regions.

## DISCUSSION

In this article, we proposed score test statistics for genome-wide screening of the extended homozygosity of the human genome. We considered three testing cases: EGHST, HMMST, and EHHST. Intuitively, EGHST might provide high values as long as either HWE or linkage equilibrium was invalid, HMMST could do so if either HWE was invalid or there existed higher order LD interaction than pairwise ones among SNPs, and EHHST might provide high scores only when the haplotype version of HWE was invalid in a chromosome region. Roughly speaking, EGHST and EHHST were two extremes: EGHST was the most aggressive one that might give many positive signals, given the high density of SNP data; hence, the presence of LD was a fact of ubiquity. EHHST was the most conservative one as it might give high scores in the presence of excess homozygosity. We started from a measure  $T$  of the extent of homozygosity, and then provided the distribution of  $T$  and its mean and variance under the null hypothesis of each test case. This facilitated the calculation of our test statistics.

By simulating data under the null hypothesis of the EGHST, we evaluated the robustness of the three tests through type I error calculations and confirmed that the EHHST was the most robust (Supplementary I). We then used coalescent programs SelSim and ms to simulate data under the neutral model. We showed that EHHST led to appropriate false-positive rates and it was robust in the presence of simple demographic population history. By comparing with the results reported in Hanchard *et al.*,<sup>26</sup> we showed that the EHHST had higher or similar power as the existing popular methods. One might want to notice that the existing popular tests usually did not follow a distribution. The EHHST, however, is asymptotically normal, which makes analysis and applications easier. We applied the tests to Hapmap Phase II data for genome-wide screening, for comparison with previously reported candidate regions, and to search for new candidate regions on the basis of high EHHST values and population differentiations. It was encouraging that our EHHST values confirmed 17 regions of excess homozygosity out of 20 candidates reported by Sabeti *et al.*<sup>36</sup> The statistics also validated the relative demographic history of African, European, and East Asian populations. Our plots suggested multiple regions of excess homozygosity. Given our ignorance about the function of many genes, it would take a long time to sort through these hints.

In summary, the main contributions are the following: we showed that the EHHST could be used to detect regions of excess homozygosity, which could be candidates of recent selection for further investigations by additional requirement, such as the criteria used in Sabeti *et al.*,<sup>36</sup> namely, selected alleles were newly arisen, were likely to be highly differentiated among populations, and had biological effects. The EHHST was conservative and robust. Compared with the existing popular methods, EHHST performed just as well or even better. Moreover, EHHST was straightforward and was asymptotically normal. In addition to EHHST, we showed that EGHST and HMMST were useful in genome-wide scans for a general picture of the strength of LD and violation of HWE by comparing test scores of different population samples. For candidate regions that had selection signals, the comparison of the three test scores might provide clues of either LD or violation of HWE or both, which lead to high test scores.

Because of the conservative nature of EHHST, one might miss some candidate regions in which HWE is roughly valid, but LD exists. Thus,

high EHHST values were not a sufficient and necessary condition for detection of selection signal. Notwithstanding, EHHST could be a new tool in addition to existing methods of detecting selection. Population geneticists have proposed several tests for inferring a selective sweep. Jensen *et al.*<sup>44</sup> summarized the most important tests, including ones based on increased LD.<sup>4,45</sup> We liked the current statistics because they exploited dense SNP genotyping and depended on minimal assumptions. Of course, the lack of a detailed model had its disadvantages. For example, our tests said nothing about the age of a favorable mutation. This issue was obviously intertwined with variations in recombination rates across the genome.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

Dr Gill introduced us to the topic of positive selection and stimulated our interest in the research. We thank Dr Pritchard's group for sharing their iHS C++ codes, which facilitated our research. This research was supported by the NIH Grant R01CA133996 from the MD Anderson Cancer Center for Fan R, and the NIH grants R01GM53275/R01MH59490 from UCLA for Lange K.

- 1 Slatkin M: Disequilibrium mapping of a quantitative-trait locus in an expanding population. *Am J Hum Genet* 1999; **64**: 1765–1773.
- 2 Bamshad M, Wooding SP: Signatures of natural selection in the human genome. *Nat Rev Genet* 2003; **4**: 99–111.
- 3 Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ: Evidence for positive selection in the Superoxide Dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* 1994; **136**: 1329–1340.
- 4 Kim Y, Nielsen R: Linkage disequilibrium as a signature of selective sweeps. *Genetics* 2004; **167**: 1513–1524.
- 5 Ronald J, Akey JM: Genome-wide scans for loci under selection in humans. *Hum Genomics* 2005; **2**: 113–125.
- 6 Vallender EJ, Lahn BT: Positive selection on the human genome. *Hum Mol Genet* 2004; **13** (Spec. No. 2): R245–R254.
- 7 Maynard SJ, Haigh J: The hitch-hiking effect of a favorable gene. *Genet Res* 1974; **23**: 23–25.
- 8 Kaplan NL, Hudson RR, Langley CH: The "hitchhiking effect" revisited. *Genetics* 1989; **123**: 887–899.
- 9 Stephan W, Wiehe THE, Lenz MW: The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor Popul Biol* 1992; **41**: 237–254.
- 10 Ewens WJ: *Mathematical Population Genetics. I. Theoretical Introduction*. New York, USA: Springer, 2004.
- 11 Fay JC, Wu CI: Hitchhiking under positive Darwinian selection. *Genetics* 2000; **155**: 1405–1413.
- 12 Fu YX, Li WH: Statistical tests of neutrality of mutations. *Genetics* 1993; **133**: 693–709.
- 13 Hudson RR, Kreitman K, Aguadé M: A test of neutral molecular evolution based on nucleotide data. *Genetics* 1987; **120**: 831–840.
- 14 Kreitman M: Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet* 2000; **1**: 539–559.
- 15 Nielsen R: Molecular signatures of natural selection. *Annu Rev Genet* 2005; **39**: 197–218.
- 16 Sabeti PC, Schaffner SF, Fry B *et al*: Positive natural selection in the human lineage. *Science* 2006; **312**: 1614–1620.
- 17 Tajima F: Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
- 18 Teshima KM, Coop G, Przeworski M: How reliable are empirical genomic scans for selective sweeps? *Genome Res* 2006; **16**: 702–712.
- 19 Watterson GA: On the number of segregating sites in genetic models without recombination. *Theor Popul Biol* 1975; **7**: 256–276.
- 20 Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: Localizing recent adaptive evolution in the human genome. *PLoS Genet* 2007; **3**: e90.
- 21 Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
- 22 Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M: A method for detecting recent selection in the human genome from allele age estimates. *Genetics* 2003; **165**: 287–297.
- 23 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **4**: e72.
- 24 Wang ET, Kodama G, Baldi P, Moyzis RK: Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci USA* 2006; **103**: 135–140.

- 25 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: Genomic scans for selective sweeps using SNP data. *Genome Res* 2005; **15**: 1566–1575.
- 26 Hanchard NA, Rockett KA, Spencer C *et al*: Screening for recently selected alleles by analysis of human haplotype similarity. *Am J Hum Genet* 2006; **78**: 153–159.
- 27 Sabeti PC: The case for selection at CCR5-Delta32. *PLoS Biol* 2005; **3**: e378.
- 28 Tang K, Thornton KR, Stoneking M: A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* 2007; **5**: e171.
- 29 McVean G: The structure of linkage disequilibrium around a selective sweep. *Genetics* 2007; **175**: 1395–1406.
- 30 Gibson J, Morton NE, Collins A: Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* 2006; **15**: 789–795.
- 31 Ayers KL, Lange K: Penalized estimation of haplotype frequencies. *Bioinformatics* 2008; **24**: 1596–1602.
- 32 Long JC, Williams RC, Urbanek M: An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 1995; **56**: 799–810.
- 33 Spencer CCA, Coop G: SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* 2004; **20**: 3673–3675.
- 34 Hudson RR: Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 2002; **18**: 337–338.
- 35 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
- 36 Sabeti PC, Varilly P, Fry B *et al*: Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007; **449**: 913–918.
- 37 Lamason RL, Mohideen MPK, Mest JR *et al*: SLC24A5, a putative cation exchanger, affects pigmentation in Zebrafish and humans. *Science* 2005; **310**: 1782–1786.
- 38 Bersaglieri T, Sabeti PC, Patterson N *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **74**: 1111–1120.
- 39 Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I: Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 2002; **30**: 233–237.
- 40 Poulter M: The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann Hum Genet* 2003; **67**: 298–311.
- 41 Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signature of natural selection. *Genome Res* 2002; **12**: 1805–1814.
- 42 Akey JM, Eberle MA, Rieder MJ *et al*: Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* 2004; **2**: e286.
- 43 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
- 44 Jensen JD, Kim Y, Dumont VB, Aquadro CF, Bustamante CD: Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* 2005; **170**: 1401–1410.
- 45 Przeworski M: The signature of positive selection at randomly chosen loci. *Genetics* 2002; **160**: 1179–1189.

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