Bioinformatics Advance Access published July 15, 2006

A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations

Chun Zhang[1§], Dione K. Bailey[1], Tarif Awad[1], Guoying Liu[2], Guoliang Xing[1], Manqiu Cao[1§§], Venu Valmeekam[2], Jacques Retief[1§§§], Hajime Matsuzaki[1], Margaret Taub[3], Mark Seielstad[4], and Giulia C. Kennedy[1*]

[1]Affymetrix Inc., 3380 Central Expressway, Santa Clara CA 95051, [2]Affymetrix Inc., 6550 Vallejo Street, Emeryville, CA 94608, [3] Department of Statistics, University of California, Berkeley, CA, [4] Department of Population Genetics and Genetic Epidemiology, Genome Institute of Singapore, Singapore 138672

Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: The identification of signatures of positive selection can provide important insights into recent evolutionary history in human populations. Current methods mostly rely on allele frequency determination or focus on one or a small number of candidate chromosomal regions per study. With the availability of large scale genotype data, efficient approaches for an unbiased whole genome scan are becoming necessary.

Methods: We have developed a new method, the whole genome long-range haplotype test (WGLRH), which uses genome-wide distributions to test for recent positive selection. Adapted from the long-range haplotype (LRH) test (Sabeti *et al.,* 2001), the WGLRH test uses patterns of linkage disequilibrium (LD) to identify regions with extremely low historic recombination. Common haplotypes with significantly longer than expected ranges of LD given their frequencies are identified as putative signatures of recent positive selection. In addition, we have also determined the ancestral alleles of SNPs by genotyping chimpanzee and gorilla DNA, and have identified SNPs where the non-ancestral alleles have risen to extremely high frequencies in human populations, termed "flipped SNPs". Combining the haplotype test and the flipped SNPs determination, the WGLRH test serves as an unbiased genome-wide screen for regions under putative selection, and is potentially applicable to the study of other human populations.

Results: Using WGLRH and high-density oligonucleotide arrays interrogating 116,204 SNPs, we rapidly identified putative regions of positive selection in three populations (Asian, Caucasian, African-American), and extended these observations to a fourth population, Yoruba, with data obtained from the International HapMap consortium. We mapped significant regions to annotated genes. While some regions overlap with genes previously suggested to be under positive selection, many of the genes have not been previously implicated in natural selection and offer intriguing possibilities for further study.

Availability: the programs for the WGLRH algorithm are freely available and can be downloaded at http://www.affymetrix.com/support/supplement/WGLRH_program.zip

Contact: Giulia_Kennedy@affymetrix.com

Supplementary Material: 3 Supplementary Tables

1 INTRODUCTION

Many phenotypic differences within and between populations may have arisen through relatively recent adaptations to local changes in the environment, and may be attributed to a subset of the genetic variation in the genome. It has been suggested that many aspects of human biology are manifestations of varying levels of positive selection. Therefore, the identification of genes that have undergone positive selection is an important step in understanding how human populations have adapted to environmental changes (Vallender and Lahn 2004).

Methods for identifying genes under positive selection have relied on molecular information and allele frequency determinations in populations around the world, first at the protein level, and later at the nucleotide sequence level. For example, a study comparing the amino-acid substitution in a coding region of a gene between human and chimpanzees has indicated that human leukocyte antigen (HLA) loci have undergone positive selection (Salamon *et al.,* 1999). In other studies, analyses of allele frequencies among human populations have also identified genes subject to positive selection, such as cytochrome P450 1A2 (CYP1A2, Wooding *et al.,* 2002). More examples of genes that may be under positive selection in specific populations have emerged (reviewed in Bamshad and Wooding 2003), but until recently, only a handful of genes had been identified in this way.

As increasing amounts of polymorphism (primarily SNP) data have become available, genome-wide methods are emerging as an attractive approach for identifying positively selected genes without *a priori* knowledge of their existence. One approach has been to look for regions with excess of rare alleles or more than expected differentiation among populations (a high F_{ST}) (Bamshad and Wooding 2003; Akey *et al.,* 2002; Bersaglieri *et al.,* 2004; Bowcock *et al.,* 1991; Hamblin *et al.,* 2002; Lewontin and Krakauer 1973). While F_{ST} might be useful to detect a signal of recent positive selection, substantial heterogeneity of F_{ST} has been observed along the human genome (Weir *et al.*, 2005). Recently, a potentially powerful long-range haplotype (LRH) test has been developed and applied to the study of specific populations in order to detect recent positive selection in a candidate chromosomal region *(*Sabeti *et al.,* 2002*)*. This method identifies common haplotypes with linkage disequilibrium (LD) ranges that are markedly longer than would be expected from their frequencies, as signatures of recent positive selection. The underlying assumption is that positive selection causes a rapid rise in allele frequency in such a short span of time that the surrounding LD has not decayed substantially (Sabeti *et al.,* 2002). The LRH test is useful for testing a single haplotype block at a time; however, due to the requirement

for a large number of simulations, it is impractical for testing the entire genome. The International

*To whom correspondence should be addressed.

§ Current address: Roche Palo Alto LLC, 3431 Hillview Avenue, Palo Alto, CA 94304 §§ Current address: Intel Corporation, 3065 Bower Avenue, Santa Clara, CA 95054 §§§Current address: Iconix Pharmaceuticals Inc., 325 E. Middlefield Rd., Mountain View, CA 94043

© 2006 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

HapMap consortium has recently identified signatures of natural selection in the human genome by detecting outliers to the genome-wide distribution for the LRH test (Altshuler *et al.,* 2005), but the methods were not clearly formulated in this paper. Here we introduce a modification of LRH, called the whole genome long range haplotype (WGLRH) test, which circumvents computationally intensive simulations by using distributions estimated from genome-wide genotype, allele frequency and haplotype data, and by testing for deviations from neutrality.

In this study, samples from Asian, Caucasian and African-American populations were rapidly analyzed using the GeneChip Mapping 100K high-density SNP oligonucleotide arrays. Genotypes on >100,000 SNPs were obtained for these populations; haplotypes were reconstructed and tested using the genome-wide distribution of the surrounding LD. In parallel, nearly 70,000 SNPs were assigned ancestral alleles by genotyping chimpanzee and gorilla, allowing for the identification of "flipped" SNPs in which the derived allele has risen to very high frequencies, another potential hallmark of natural selection (Wooding *et al.,* 2002; Kennedy *et al.,* 2003). We mapped a subset of the putative regions, identified by high significance in the WGLRH test, to annotated genes and placed these genes into functional categories using the Gene Ontology (GO) database. Finally, we extended our analytical methods to 42 Yoruba samples obtained from the HapMap consortium (http://www.hapmap.org/) and compared the results amongst these four populations.

2 ALGORITHM

The workflow of our WGLRH algorithm is as follows: first, core regions (haplotype blocks with extremely low historic recombination) were identified throughout the genome. Second, we estimated the haplotypes and haplotype frequencies in these regions, and 500 kilobase extended regions adjacent to the core regions. Third, extended haplotype homozygosity (EHH) and relative extended haplotype homozygosity (REHH) were calculated in the extended regions based on the estimated haplotypes and haplotype frequencies. We then tested each haplotype in the core regions (core haplotype) against the genome-wide distribution of REHH for the haplotypes with similar frequencies. This allowed us to identify core haplotypes with LD ranges that were unusually longer than would be expected from their frequencies. Finally, to reduce the false positive rate and also improve statistical efficiency, we only considered the core regions where the adjacent 500 kilobase extended regions overlapped with "flipped" SNPs (SNPs having derived alleles that have risen to extremely high frequencies), as regions that are likely to have undergone recent positive selection.

2.1 Identification of core regions

We define a core region as a region that simultaneously meets two definitions of haplotype blocks: Gabriel's D' (Gabriel *et al.,* 2002) and the four-gamete test (Wang *et al.,* 2002). Gabriel's D' uses the lower bound of 95% confidence interval of D'>0.7 and the upper bound of 95% confidence interval of D'>0.98. Meanwhile, the four-gamete test identifies haplotype blocks based on the distribution of observed crossovers between loci: recombination must have occurred if all four gametes are observed between a pair of di-allelic loci; if we observe no more than three gametes, there is no evidence of recombination. We used both definitions of haplotype blocks to define the core regions, so that recombination in the core regions will be extremely rare.

2.2 Estimation of haplotype and haplotype frequency

Haploview (Barrett *et al.,* 2005) was used to estimate the haplotypes and the haplotype frequencies in the identified core regions. To infer the range of LD surrounding the haplotypes in the core regions, we also examined the 500 kilobase extended regions adjacent to the core regions, and estimated haplotypes and haplotype frequencies for the entire interval from the core region to the extended regions. Only common haplotypes (frequency >5%) were retained in this study.

2.3 EHH and REHH

Extended haplotype homozygosity (EHH) is a probability that two randomly chosen haplotypes carrying the candidate core haplotype are homozygous for the entire interval spanning the core region to locus x (Sabeti *et al.,* 2002). Suppose we have one core region with i=1,2,…,I core haplotypes and the entire region (core region + extended region) with j=1,2,…,J haplotypes. J>=I. Denote $EHH_x^{(i)}$ as the EHH for the haplotypes ranging from the core region to locus x in the extended 500 kilobase regions, carrying core haplotype i. We first grouped the haplotypes for the entire region by the core haplotypes they carry. For the group carrying core haplotype i, we then combined the identical haplotypes from the core region to locus x and calculated the frequency of each unique haplotype to locus x. Let $f_j^{(i)}$ denote the frequency of the haplotype j in group i, and $H_x^{(ik)}$ denote the kth set of haplotypes that are identical from the core region to locus x in group i, k=1,2,…,K. Then the frequency of the kth unique haplotype from the core region to locus x in group i can be calculated as:

$$f_x^{(ik)} = \frac{\sum_{j \in H_x^{(ik)}} f_j^{(i)}}{\sum_{j=1}^{J} f_j^{(i)}}$$

for x=1,2,…,X; i=1,2,…,I; k=1,2,…,K. Then $EHH_x^{(i)}$ can be calculated by using the following:

$$EHH_x^{(i)} = \sum_k \left(f_x^{(ik)}\right)^2$$

for x=1,2,…,X; i=1,2,…,I. In the case of K=1, let EHH=1.

Many studies have indicated that various chromosomal regions could have different recombination rates. The LD would be stronger in recombination "cold spots" than in recombination "hot spots" which raises the possibility that a greater measure of LD could be due to low recombination rates in a particular region, rather than recent positive selection. Taking this into consideration, we calculated relative extended haplotype homozygosity (REHH) as a measure of LD surrounding a haplotype of interest to minimize the regional recombination effects.

We define REHH as the extended haplotype homozygosity of one core haplotype relative to other core haplotypes in a core region, adjusting for the frequency of the core haplotypes. Let $REHH_x^{(i)}$ denote the REHH of core haplotype i at locus x; it can be calculated as:

$$REHH_x^{(i)} = \frac{f_i EHH_x^{(i)}}{\sum_{l \neq i} f_l EHH_x^{(l)}}$$

Here f_i denotes the frequency of core haplotype i. According to this definition, the core haplotypes with longer range of LD and higher frequency would result in greater REHH values.

To make the REHH for different chromosomal regions comparable, we calculate REHH at every 50 kilobase in the 500 kilobase extended regions (x=1,2,…,10 indicating 50 kb, 100 kb,…,500 kb from the core region) of the core haplotypes. For an extended region where there is no marker at locus x, the closest neighbor in the interval of [50 kb*(x-1), 50 kb*x] was used for calculating the REHH at locus x instead. When the extended region is so sparse that we cannot find a marker in the interval of [50 kb*(x-1), 50 kb*x], the REHH at locus x is recorded as missing. The core haplotype with maximum summation of REHH over all locus (x=1,2,…,10) was chosen to represent the corresponding core region.

2.4 Hypothesis testing for each core haplotype

The core haplotypes that experienced recent positive selection should have higher REHH compared to the neutral core haplotypes on the whole genome with a similar frequency. By assuming the majority of markers in the human genome are neutral for autosomal chromosomes, we used the genome-wide distribution of REHH for the core haplotypes with similar frequencies to a core haplotype of interest ($|f_j - f_i| < 0.10$, f_i: the frequency of the haplotype of interest, f_j: the frequency of the core haplotypes with similar frequency) as the distribution of REHH under neutrality. For each locus x in the extended region of one core haplotype, we assume that the REHH under neutrality follows a gamma distribution, with the shape and rate parameters estimated from the REHH for the haplotypes with <10% frequency difference by using maximum likelihood methods. The REHH for the core haplotype of interest will then be tested against the gamma distribution, and a low p-value would suggest a REHH higher than expected under neutrality given the frequency of the haplotype. This indicates a putative signature of recent positive selection.

Since the X chromosome may be more exposed to natural selection due to its haploid state in males and smaller effective population size, we used genome-wide common haplotypes complying with Hardy Weinberg Equilibrium (HWE) to infer the null distribution of REHH for the X chromosome.

2.5 Identification of flipped SNPs

Chimpanzee and gorilla nucleotide sequence differs from human by only 1.5% and 2.1%, respectively (Britten 2002; Goodman *et al.,* 1998). Thus, it is possible to derive accurate genotype information on non-human primates using human SNP arrays, and therefore assign ancestral allele status (Hacia *et al.,* 1999; Kennedy *et al.,* 2003). We genotyped one chimpanzee and one gorilla on the Affymetrix GeneChip Mapping 100K arrays and only SNPs that were homozygous and identical in both the chimpanzee and gorilla were assigned as ancestral (Kennedy *et al.,* 2003). We computed the allele frequencies for each SNP and identified those SNPs in which the ancestral allele frequency was <0.15, or conversely the non-ancestral allele frequency was >0.85, in one or more populations. These SNPs were designated as "flipped SNPs". Under infinite sites model, a recombination event should have occurred in a 2-SNP core region with a core haplotype composed of two derived alleles. Therefore, such core regions were removed from the study. The remaining core regions where the adjacent 500 kilobase regions overlapped with the flipped SNPs were selected for a final screen.

2.6 Multiple testing corrections

To control for false positives, the false discovery rate (FDR) approach was used for multiple testing corrections (Benjamini and Hochberg 1995). Let $P_{(1)}{}^x < P_{(2)}{}^x < \ldots < P_{(m)}{}^x$ denote the ordered p-values of the final m core haplotypes at locus x in the 500 kilobase regions, the threshold is determined by $T_{BH}{}^x = \max\left\{P_{(i)}{}^x : P_{(i)}{}^x \le \alpha \frac{i}{m}, 0 \le i \le m\right\}$. All the tests at locus x with p-value $< T_{BH}{}^x$ are rejected at α =0.05. We define the core haplotypes with rejected tests at any locus x in the 500 kilobase region as significant haplotypes. Some of the haplotypes could be rejected at more than one locus. For these haplotypes, we report all the p-values associated with the rejected tests and the loci where these p-values were obtained.

3 IMPLEMENTATION

3.1 Samples

Affymetrix GeneChip® Mapping 100K arrays were used to genotype 116,204 SNPs in three populations (37 Asian, 42 Caucasian, and 42 African-American individuals). The data are publicly available and have been described elsewhere (Matsuzaki *et al.,* 2004). Only SNPs with known physical locations mapped onto the National Center for Biotechnology Information (NCBI) genome build 34 were included in this study (115,571 SNPs). We also selected 42 unrelated individuals from the Yoruba HapMap genotype data (**http://www.hapmap.org/**). To make the results of the Yoruba data comparable to the Affymetrix 100K data, we extracted 102,360 SNPs that directly overlap with GeneChip Mapping 100K SNPs from the 1,064,173 Yoruba HapMap SNPs, which results in approximately 90% of the Affymetrix 100K SNPs (115,571) in the Yoruba genotype data.

3.2 Detecting regions of recent positive selection in the human genome

We identified a large number of core regions throughout the genome (11,308 for Asians, 10,136 for African-Americans, 12,145 for Caucasians and 8,090 for Yoruba). Figure 1 shows the number of core regions we found in the 23 chromosomes. In all populations studied, we found statistically significant longer core regions on the X chromosome than on the autosomes (Asian: mean 15.736 kb for autosomes, 28.807 for X; Caucasian: 15.515 for autosomes, 33.454 for X; African-American: 8.610 for autosomes, 19.713 for X; Yoruba: 9.437 for autosomes, 38.857 for X). This is consistent with the prediction that LD is greater, and the size of regions with a single genetic history is larger, on the X chromosome (Schaffner 2004). Yoruba and African-American samples show both fewer core regions, as well as regions of significantly shorter ranges on the autosomes. This finding is consistent with previous studies showing that African populations have a lower level of LD and smaller haplotype blocks (Tishkoff and Verrelli 2003).

We assigned ancestral states to 69,500 SNPs (35446 SNPS with ancestral allele A, 34054 SNPs with ancestral allele B) on the

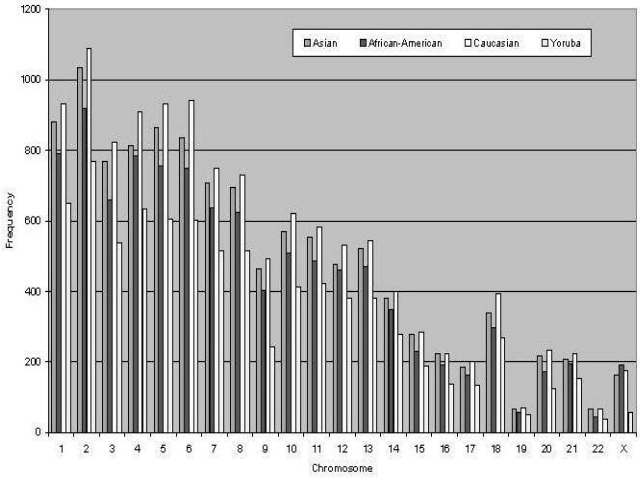

Fig. 1. Number of core regions identified in four populations (Asian, African-American, Caucasian and Yoruba) in each of the 23 chromosomes.

Mapping 100K arrays, of which 62,895 SNPs (32041 SNPs with ancestral allele A, 30854 SNPs with ancestral allele B) overlapped with the Yoruba HapMap data (Supplemental Table 1). Table 1 shows the distribution of the ancestral allele frequencies for each population. As expected from theoretical predictions (Watterson and Guess 1977), for the vast majority of SNPs, the ancestral allele is observed at a higher frequency than the non-ancestral allele. However, for 2-5% of the SNPs, termed "flipped SNPs" the derived allele has achieved high frequency (>0.85), possibly because of either positive or negative selection (Wooding *et al.,* 2002). We found 5,322 flipped SNPs in the Asian population (2616 from ancestral allele A, 2706 from allele B), 2,127 in the African-American population (1084 from allele A, 1043 from allele B), 4,850 in the Caucasian population (2403 from allele A, 2447 from allele B), and 2,786 flipped SNPs in the Yoruba population (1734 from allele A, 1052 from allele B) (Table 1).

The flipped SNPs were used to screen for the core regions that are likely to have undergone positive selection. 56 of the Asian core regions (0.5%), 53 of the Caucasian core regions (0.4%) and 18 African-American core regions (0.2%) are 2-SNP core regions with a haplotype composed of two derived alleles, and hence were removed from the study. Among the remaining core regions, 6935 (61%) of the Asian core regions, 3484 (36%) of the African-American core regions, 7202 (59%) Caucasian regions and 3582 (44%) Yoruba core regions overlapped with the flipped SNPs within the adjacent 500 kilabase intervals. From these core regions, we found 34, 37, 23 and 32 significant regions (at significance level of 0.05 after FDR correction) from the Asian, Caucasian, African-American and Yoruba samples, respectively (Supplemental Table 2). A summary of the significant core haplotypes can be found in Table 2. We found no significant differences in the number of significant core regions detected on the X chromosomes relative to the autosomes. This suggests that the majority of effects seen on the X chromosome relative to the autosomes may reflect pronounced sensitivity to genetic drift due to the smaller effective population size of the X chromosome.

Figure 2 shows the distribution of the REHH for the haplotypes with <10% frequency differences with one selected significant haplotype in each population (Supplemental Table 2). It demonstrates that the gamma distribution with estimated shape and rate parameters well represents the distribution of REHH under neutrality and reflects the deviation of the REHH for the targeted haplotype from the null distribution. For the REHH sets that represent the null distribution, no severe violations to the gamma distribution were observed in our data. Figure 3 shows a significant core region (haplotype 2: significant REHH at 200 kb (p-value $1.80\text{x}10^{-7}$, FDR corrected p-value $1.50\text{x}10^{-4}$), 250 kb (p-value $2.46\text{x}10^{-7}$, corrected $3.10\text{x}10^{-4}$), 300 kb (p-value $2.13\text{x}10^{-6}$, corrected $1.93\text{x}10^{-3}$), 350 kb (p-value $9.11\text{x}10^{-9}$, corrected $6.21\text{x}10^{-6}$) and 400 kb (p-value $3.92\text{x}10^{-8}$, corrected $2.13\text{x}10^{-5}$) from the core region.) on chromosome 6 (SNP_A-1669684, SNP_A-1669802) in the Asian population. Two haplotypes with equivalent frequency (0.5) were estimated in this core region. While for both haplotypes the EHH values decrease with the distance from the core region, haplotype 2 has consistently higher EHH values than haplotype 1. The REHH values on the extreme of the null distribution (represented by the boxplots at every 50kb) indicate unusually long LD given the frequency of the haplotype. One SNP (SNP_A-1693376) which is 157 kb away from the core region was flipped only in the Asian population (non-ancestral allele frequency 0.57, 0.87, 0.5, 0.6 in African-American, Asian, Caucasian and Yoruba respectively). In this example, the haplotype test and the flipped SNP approaches were consistent in identifying this region as a signature of positive selection in the Asian population.

Table 1. Summary of ancestral alleles

Population	Ancestral Allele	Mean Allele A freq	Stdev of Allele A freq	Number of Flipped SNPs	Number of SNPs Assigned Ancestral Allele
African American	A	0.718	0.239	1084	35446
Asian	A	0.688	0.292	2616	35446
Caucasian	A	0.686	0.281	2403	35446
Yoruba	A	0.712	0.268	1734	32041
African American	B	0.290	0.242	1043	34054
Asian	B	0.322	0.296	2706	34054
Caucasian	B	0.325	0.285	2447	34054
Yoruba	B	0.268	0.254	1052	30854

Freq: frequency; Stdev: standard deviation.

Table 2. Summary of significant core haplotypes.

Population	SZ	# significant core haplotypes	SNP number		Size (bp)	
			Min-Max	Med	Mean	Stdev
Asian	37	34	2-7	2	10,515	14,764
African-American	42	23	2-4	2	6,203	11,570
Caucasian	42	37	2-5	2	13,710	30,707
Yoruba	42	32	2-6	2	4,184	7,787

SZ: sample size; Min-Max: minimum to maximum, indicates the range of SNP number in the significant core haplotypes. Med: median; Stdev: standard deviation of the size of the significant core haplotypes.

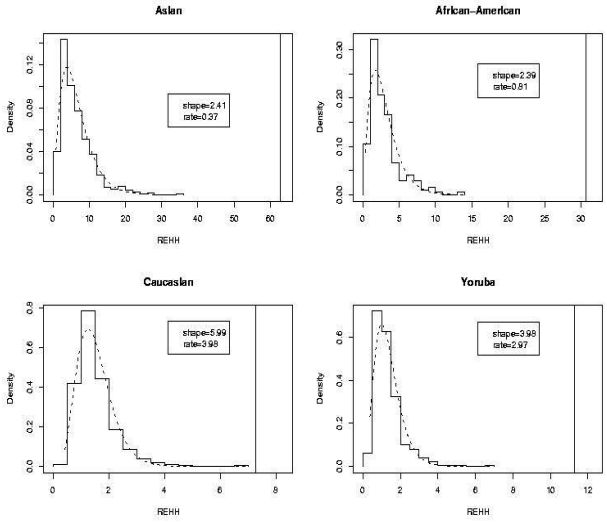

Fig. 2. Distribution of REHH for the one significant haplotype in each population. Distribution of the REHH for the haplotypes with similar frequency to the haplotypes of interest (<10% frequency difference) on one significant region in Asian (chromosome 1: 103027364—103029079, p-value: $5.27x10^{-9}$), African-American (chromosome 12: 78057101—78058473, p-value: $1.22x10^{-9}$), Caucasian (chromosome 9: 28841909—28844830, p-value: $5.45x10^{-8}$) and Yoruba (chromosome 10: 119820320—119820337, p-value: $1.89x10^{-11}$) samples respectively. The dotted curve stands for the fitted gamma density function with the shape and rate parameters estimated from the data. The vertical line represents the REHH at one of the significant loci for this haplotype (Asian: 62.704 at 200 kb; African-American: 30.711 at 500 kb; Caucasian: 7.281 at 100 kb; Yoruba: 11.292 at 350 kb).

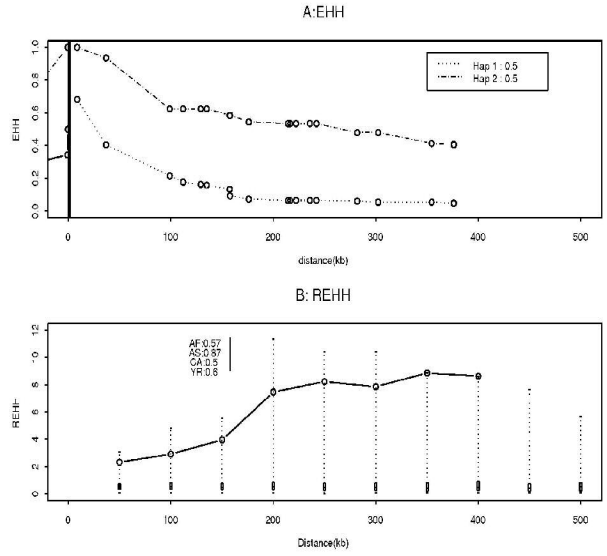

Fig. 3. EHH and REHH for one core region. Panel A demonstrates the distribution of EHH on the right of one core region on chromosome 6 in Asian population. The vertical black line represents the core region, the X axis denotes the distance from the core region and the y axis denotes the value of EHH. Dotted lines represent the two haplotypes with the same frequency (50%). Panel B demonstrates the distribution of REHH for haplotype 2 on the right of the core region. The X axis denotes the distance from the core region, and the Y axis denotes the value of REHH. The boxplots on the bottom represent the null distribution of REHH at each location. The vertical line on the upper left indicates the position of one flipped SNP (SNP_A-1693376), and the frequencies of the non-ancestral allele in Asian, African-American, Caucasian and Yoruba are shown on the plot (0.57, 0.87, 0.5, 0.6 in African-American, Asian, Caucasian and Yoruba respectively).

3.3 Mapping positively selected regions to genome annotations

We also mapped the significant regions to genome annotations. RefSeq annotations from NCBI build 34 of the human genome were downloaded from the UCSC genome browser (http://genome.ucsc.edu) and annotation information on the 100K Mapping array set was obtained from NetAffx™ (http://www.affymetrix.com). The genomic intervals of the haplotypes were then searched against the RefSeq annotations to identify transcripts that have at least 1 bp overlap with a WGLRH region.

A subset of the identified WGLRH regions mapped to annotated transcripts in the genome (Supplemental Table 3). These transcripts belong to cell adhesion, regulation of transcripts, transcport/ion transport, protein/ubiquitin processing and other functional categories. Many of these genes have not been previously implicated in natural selection and offer intriguing possibilities for further study. However, some regions overlap with genes previously suggested to be under selection. For example, *Eyes absent homolog 4* (EYA4), a member of the vertebrate *Eya* family of transcriptional activators, has been indicated to be a signature of human-specific positive selection (Clark *et al.,* 2003). In other studies integrin alpha 8 (ITGA8, Akey *et al.,* 2004) and MAX interactor 1 (MXI1, Storz *et al.,* 2004) have also been identified as genes that have undergone recent positive selection.

4 DISCUSSION

In summary, we have described a rapid set of approaches to screen human populations for regions of selection: we first generated highly accurate genotype information on 116,204 genome-wide SNPs using GeneChip Mapping 100K arrays. The samples in this study were genotyped in a matter of days. Second, we adapted the LRH test of Sabeti *et al.* to enable a genome-wide test for selection. Third, we identified thousands of derived alleles with extremely high frequency relative to the ancestral allele, another potential hallmark of selection. Finally, we mapped putative regions under selection, defined as core regions that are statistically significant from the WGLRH tests, to annotated genes. Many of these genes play important roles in human health and can now serve as starting points for formulating biological hypotheses.

Tests of selection traditionally use coalescent simulations under a model of demographic history. This poses two problems: first, the approach can be computationally intensive, and second, the models tend to oversimplify demographic history. Alternatively, the use of genome-wide haplotypes to infer the distribution of linkage disequilibrium under neutrality, provides a realistic and computationally efficient solution for the identification of recent positive selection. While it might be possible that the unusually long LD intervals are due to recombination "cold spots", we have minimized this possibility by using the relative haplotype homozygosity (REHH), a measure of the LD surrounding a haplotype compared to other haplotypes in the same core region, adjusting for their haplotype frequencies. It is important to note that the density of the markers might affect the patterns of LD, hence have an impact on the distribution of REHH. While how the distribution of REHH would depend on the density of markers is not clear, a denser map of markers could potentially enable more

detailed investigation of recent positive selection on the human genome.

The identification of flipped SNPs by genotyping chimpanzee and gorilla provides a complementary approach to the haplotype test for the identification of regions under selection. It is important to use ancestral information to unambiguously identify the derived states of the SNPs. On the other hand, demographic factors independent of selection, such as population bottlenecks and genetic drift, are likely to have led to the flipping of SNPs in some instances and populations. Therefore, by combining the flipped SNPs and the haplotype test, we filter out the core regions that are unlikely to be under selection, and thereby reduce the number of required tests and increase power.

Finally, we have provided a model-free, computationally efficient solution for the identification of positive selection in the whole genome using high-density microarrays with 116,204 SNPs. Our approach is applicable to the study of other human populations where new candidates for natural selection can be rapidly identified, and will hopefully facilitate our understanding of human evolutionary history and the molecular basis of complex human diseases. Advances in genotyping technology, which enable the mapping of > 500,000 SNPs, will further enhance the ability to find regions of positive selection using this method.

ACKNOWLEDGEMENTS

We thank two reviewers for insightful comments and Dr. Yongchao Ge for valuable discussions.

REFERENCES

Akey,J.M., Eberle,M.A., Rieder,M.J., Carlson,C.S., Shriver,M.D., Nickerson,D.A. and Kruglyak,L. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PloS Biology,* **2**(10), 1591-1599.

Akey,J.M., Zhang,G., Zhang,K., Jin,L. and Shriver,M.D. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research,* **12**, 1805-1814.

Altshuler,D., Brooks,L.D., Chakravarti,A., Collins,F.S., Daly,M.J. and Donnelly,P. (2005) International HapMap Consortium. A haplotype map of the human genome. *Nature*, **437**, 1229-1320.

Bamshad,M. and Wooding,S.P. (2003) Signatures of natural selection in the human genome. *Nat Rev Genet***, 4,** 99-111.

Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21,** 263-265.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser., **B*** **57**, 289-300.

Bersaglieri,T. *et al.* (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.,* **74,** 1111-1120.

Bowcock,A.M., Kidd,J.R., Mountain,J.L., Hebert,J.M., Carotenuto,L., Kidd,K.K. and Cavalli-Sforza,L.L. (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc. Natl. Acad. Sci. U S A*, **88,** 839-843.

Britten,R.J. (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl. Acad. Sci. U S A*, **99,** 13633-13635.

Clark,A.G. *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science,* **302**(5652), 1960-1963.

Gabriel,S.B. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296,** 2225-2229.

Goodman,M., Porter,C.A., Czelusniak,J., Page,S.L., Schneider,H., Shoshani,J., Gunnell,G. and Groves,C.P. (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.*, **9**, 585-598.

Hacia,J.G. *et al.* (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.,* **22**, 164-167.

Hamblin,M.T., Thompson,E.E., and Di Rienzo,A. (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.*, **70,** 369-383.

Lewontin,R.C. and Krakauer,J. (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics,* **74**, 175-195.

Kennedy,G.C. *et al.* (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.***, 21,** 1233-1237.

Matsuzaki,H. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109-111.

Sabeti,P.C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature,* **419**, 832-837.

Salamon,H. *et al.* (1999) Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics,* **152**, 393-400.

Schaffner,S.F. (2004) The X chromosome in population genetics. *Nat. Rev. Genet.,* **5**, 43-51.

Storz,J.F., Payseur,B.A. and Nachman,M.W. (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution*, **21**(9), 1800-1811.

Tishkoff,S.A. and Verrelli,B.C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.,* **4,** 293-340.

Vallender,E.J. and Lahn,B.T. (2004) Positive selection on the human genome. *Hum. Mol. Genet.,* **13** Spec No 2, R245-254.

Wang,N., Akey,J.M., Zhang,K., Chakraborty,R. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.,* **71**, 1227-1234.

Watterson,G.A. and Guess,H.A. (1977) Is the most frequent allele the oldest? *Theor. Popul. Biol.*, **11,** 141-160.

Weir,B.S., Cardon,L.R., Anderson,A.D., Nielsen,D.M. and Hill,W.G. (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15**(11), 1468-1476.

Wooding,S.P., Watkins,W.S., Bamshad,M.J., Dunn,D.M., Weiss,R.B. and Jorde,L.B. (2002) DNA sequence variation in a 3.7-kb noncoding sequence 5-prime of the CYP1A2 gene: implications for human population history and natural selection. *Am. J. Hum. Genet.,* **71**, 528-542.