Hum Genet (2006) 119: 92–102
DOI 10.1007/s00439-005-0090-0

ORIGINAL INVESTIGATION

**Emily C. Walsh · Pardis Sabeti · Holli B. Hutcheson
Ben Fry · Stephen F. Schaffner · Paul I. W. de Bakker
Patrick Varilly · Alejandro A. Palma · Jessica Roy
Richard Cooper · Cheryl Winkler · Yi Zeng
Guy de The · Eric S. Lander · Stephen O'Brien
David Altshuler**

Searching for signals of evolutionary selection in 168 genes related to immune function

Received: 30 May 2005 / Accepted: 6 September 2005 / Published online: 14 December 2005

Abstract Pathogens have played a substantial role in human evolution, with past infections shaping genetic variation at loci influencing immune function. We selected 168 genes known to be involved in the immune response, genotyped common single nucleotide polymorphisms across each gene in three population samples (CEPH Europeans from Utah, Han Chinese from Guangxi, and Yoruba Nigerians from Southwest Nigeria) and searched for evidence of selection based on four tests for non-neutral evolution: minor allele frequency (MAF), derived allele frequency (DAF), Fst versus heterozygosity and extended haplotype homozygosity (EHH). Six of the 168 genes show some evidence for non-neutral evolution in this initial screen, with two showing similar signals in independent data from the International HapMap Project. These analyses identify two loci involved in immune function that are candidates for having been subject to evolutionary selection, and highlight a number of analytical challenges in searching for selection in genome-wide polymorphism data.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00439-005-0090-0 and is accessible for authorized users.

Emily C. Walsh, Pardis Sabeti, Holli B. Hutcheson, and Ben Fry have contributed equally to this work and Stephen O'Brien and David Altshuler have jointly supervised this project

E. C. Walsh (✉)
Novartis Institutes for Biomedical Research, 250 Mass Ave, Cambridge, MA 02139, USA
E-mail: Emily.walsh@novartis.com
Tel.: +1-617-8713319
Fax: +1-617-8717080

E. C. Walsh · P. Sabeti · B. Fry · S. F. Schaffner
P. I. W. de Bakker · P. Varilly · A. A. Palma · J. Roy
E. S. Lander · D. Altshuler
Broad Institute of MIT and Harvard, 1 Kendall Square, Cambridge, MA 02139, USA

H. B. Hutcheson · C. Winkler · S. O'Brien
Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD, USA

R. Cooper
Department of Preventive Medicine and Epidemiology, Loyola University Medical School, Maywood, IL, USA

Y. Zeng
Institute for Viral Disease Control and Prevention, Beijing, China

G. de The
The Institute Pasteur, Paris, France

P. I. W. de Bakker · D. Altshuler
Massachusetts General Hospital, Boston, MA, USA

D. Altshuler
Harvard Medical School, Boston, MA, USA

Report

It is likely that major selective forces in human history included infection, starvation/dehydration, and surviving the peripartum period. Patterns of genetic variation can be used to infer a past history of selection at a locus, as different types of selection (purifying, positive, balancing) each leave a telltale signature in patterns of allele frequencies and linkage disequilibrium. To date, some of the strongest evidence for selection in the human genome have been identified at immunity-related loci, for example, at the HLA (Walsh et al. 2003), and loci involved in resistance to malaria (Hamblin et al. 2002; Sabeti et al. 2002; Tishkoff et al. 1996). It would be highly valuable to identify additional genes that show signs of non-neutral evolution, as these would provide clues to genes which are lynchpins of the immune

response, and in whom variation might influence infections and autoimmunity in the current population.

We set out to characterize genetic variation in 168 genes spanning 64 chromosomal regions (Table 1 and Supplemental Table 1). Genes were chosen for their broad immunological roles and/or because they lay in pathways previously implicated in HIV susceptibility (O'Brien and Nelson 2004). Our goals in generating this dataset were twofold: first, as the project was undertaken before data was available from the International Haplotype Map Project, we aimed to guide selection of "tag" SNPs for subsequent genetic association studies. Second, given that loci involved in the immune response are candidates for having been subject to non-neutral evolution, we wanted to evaluate whether any of these genes stood out from distributions (a) expected under neutral evolution based on simulations and (b) observed at unlinked loci in empirical data from the 168 genes.

Our study included three population samples: 12 multi-generational families from Utah residents with European ancestry from the CEPH collection (CEU) (96 independent chromosomes represented), 30 unrelated Han Chinese trios from Guangxi (HCG) (120 independent chromosomes) and 30 unrelated trios of the Yoruba people from the Southwest of Nigeria (YRS) (124 independent chromosomes). We surveyed variation in a region spanning 20 kb up- and down-stream of each coding sequence, with an additional ~10 SNPs per locus typed at 40 kb intervals to evaluate signals of selection based on long-range haplotypes. The SNPs were selected from public databases in multiple batches over a 18 months period from July 2002 to December 2003; preference was given to "double hit" SNPs which have been shown to be more likely to be polymorphic (Reich et al. 2003). These criteria bias ascertainment towards higher frequency variants, and thus reduce representation of rare and population-specific variation.

Overall, assays for 2,504 "gene-based" SNPs were tested, with 1,688 working, polymorphic assays developed (the remainder failing to meet quality standards or non-polymorphic). Genotyping was done using three technology platforms (Sequenom, Illumina, and ParAllele), providing an independent assessment of the comparability and performance of each method (Supplemental Fig. 1 and Materials and methods). For those interested in using the data to characterize LD in each region, or as a guide to association studies, detailed information on the data is presented in Supplemental Table 2 and on-line (http://www.broad.mit.edu/personal/pardis/CandidatePaper/). This includes assays for each SNP, genotypes in each DNA sample, characteristics of linkage disequilibrium and haplotypes, and suggested tag SNPs. These data may be used in conjunction with similar datasets collected by the International Haplotype Map Project (http://www.hapmap.org), and/or Perlegen Sciences (Hinds et al. 2005), to guide tag SNP selection at these loci.

We examined the 168 genes for possible positive directional and balancing selection events based on four statistical tests: three based on allele frequencies, and a fourth based on the extent of long-range LD. A key challenge in such an analysis is that SNP ascertainment plays a major role in determining the spectrum of alleles observed. These limitations mean (a) that statistical tests must be tailored to incorporate known ascertainment bias, and (b) that observations from such a study offer no more than a screen for loci of interest which must then be subjected to more detailed investigation. Specifically, for each test, we estimated statistical significance by comparing the observed value for each gene to a theoretical distribution generated by a population genetic simulation. The population genetic simulation uses parameters for population demography and recombination rate variation tailored to match a host of features of population genetic data in large-scale datasets, including the SNP ascertainment scheme used to generate dbSNP and select SNPs (Schaffner et al. 2005). While the full details of each test are presented in Methods, we describe each test in brief below.

The *minor allele frequency* (*MAF*) test assesses the percentage of SNPs within a locus with low MAF ($< 10\%$) and the percentage of SNPs within that locus with high MAF ($\mathrm{MAF} > 40\%$) and plots those percentages compared to simulated control (Fig. 1a–c). A selective sweep causes a disproportionate number of SNPs with rare MAFs, while recent balancing selection causes an excess of SNPs with high MAF. We selected the frequency thresholds based on their sensitivity to detect selection in all three populations using power calculations for extensive simulations of selection (Schaffner et al. 2005).

The *derived allele frequency* (*DAF*) test is similar to the MAF test however this test assesses the frequency of derived (non-ancestral) alleles as determined by comparison to chimpanzee sequence. Derived alleles provide a specific utility in detecting genetic hitchhiking; that is, variation linked to an allele under positive selection can hitchhike to low or high frequency, which could be detected by an excess of derived alleles (Fay and Wu 2000). We assess the percentage of SNPs in each locus with $\mathrm{DAF} > 80\%$ and the percentage of $\mathrm{DAF} \leq 20\%$ and plot compared to a simulated control (Fig. 1d–f).

F_{ST} versus heterozygosity (*F_{ST} versus H*) test combines two traditional methods for detecting selection. As geographically separated populations may be subject to distinct selective environments, selection can increase population differentiation at a selected locus. The F_{ST} statistic (Taylor et al. 1995) assesses this by comparing the frequency of an allele between populations. Heterozygosity assesses the genetic diversity within a population; a selective sweep can reduce genetic diversity and balancing selection can increase genetic diversity. We plot the average F_{ST} for a region between the population of study and both other populations in comparison to the average within-population haplotype heterozygosity for that region (Fig. 1g–i). By comparing F_{ST} and heterozygosity, it is possible to identify (a) loci where F_{ST} is high and heterozygosity is low, indicating a

Table 1 Region by region analysis of SNP coverage. Gene clusters are named according to the first gene in the cluster per chromosomal position and are given in italics. Average Maximal r^2 (AM r^2) is computed as the average of the maximal r^2 values for each SNP in the dataset. Fraction Needed for each region is calculated as the number of polymorphic SNPs less the number of SNPs highly correlated with other SNPs divided by the total number of SNPs. The final column indicates whether additional tag SNPs are likely necessary for that gene/region in a particular populations (C CEU, H HCG, Y YRS). Tags are listed in Supplemental Table 3

Gene symbol	CHR	Good SNPs	kb covered	Average density	CEU AM r^2	HCG AM r^2	YRS AM r^2	CEU Fraction Needed	HCG Fraction Needed	YRS Fraction Needed	More tSNPs req'd
LCK	1	3	0.65	0.22	0	0	0.03	0.667	0.667	1	CHY
ABCD3	1	11	158.13	14.38	0.58	0.59	0.68	0.727	0.091	0.636	CY
VAV3	1	40	393.75	9.84	0.59	0.48	0.41	0.875	0.75	0.9	CHY
CD58	1	15	56.45	3.76	0.62	0.58	0.37	0.333	0.533	0.867	HY
MNDA cluster	1	266	2972.23	11.17	0.55	0.55	0.47	0.741	0.699	0.797	CHY
PTPRC	1	22	118.11	5.37	0.5	0.56	0.29	0.864	0.773	0.773	CHY
AGT	1	10	11.58	1.16	0.58	0.54	0.24	1	0.9	1	CHY
MAL	2	6	28.26	4.71	0.7	0.67	0.48	0.833	0.5	0.5	CY
IL1R2 cluster	2	70	602.67	8.61	0.64	0.63	0.53	0.671	0.743	0.843	CHY
IL1A cluster	2	45	182.83	4.06	0.57	0.4	0.41	0.8	0.333	0.889	CY
CD28	2	28	31.36	1.12	0.91	0.89	0.76	0.571	0.429	0.857	CY
SLC11A1	2	6	12.08	2.01	0.47	0.46	0.22	0.833	0.833	1	CHY
NCL	2	5	9.56	1.91	0.24	0.36	0.16	1	1	1	CHY
CAV3	3	8	12.96	1.62	0.37	0.44	0.27	1	1	1	CHY
CCR9 cluster	3	24	601.44	25.06	0.63	0.77	0.53	0.75	0.667	0.875	CHY
GC	4	41	42.35	1.03	0.63	0.68	0.63	0.244	0.585	0.854	HY
IL8 cluster	4	124	561.32	4.53	0.68	0.64	0.75	0.218	0.387	0.694	
CXCL9 cluster	4	10	117.88	11.79	0.5	0.56	0.58	0.8	0.6	1	CHY
CXCL13	4	34	100.08	2.94	0.41	0.44	0.4	0.471	0.618	0.882	CHY
ABCG2	4	36	66.58	1.85	0.72	0.69	0.51	0.583	0.778	0.778	CHY
FACL6 cluster	5	25	153.27	6.13	0.76	0.76	0.57	0.6	0.36	0.72	CY
IRF1 cluster	5	75	831.84	11.09	0.76	0.69	0.59	0.56	0.413	0.827	CY
IL9	5	25	3.58	0.14	0.38	0.25	0.7	0.4	0.36	0.44	
ETF1 cluster	5	15	160.40	10.69	0.7	0.62	0.71	0.6	0.133	0.733	CY
CD14	5	10	1.45	0.15	0.99	1	0.81	0.5	0.4	0.6	
ITK	5	14	74.20	5.30	0.41	0.49	0.19	0.571	1	0.786	CHY
LCP2	5	16	49.30	3.08	0.33	0.31	0.12	0.813	0.875	0.875	CHY
FYN	6	85	212.14	2.50	0.89	0.89	0.79	0.412	0.353	0.529	
IL6	7	6	4.80	0.80	0.52	0.44	0.55	0.333	0.5	0.667	
PPIA	7	4	4.95	1.24	0.87	0.91	0.73	0.75	0.5	1	CY
ABCB1	7	26	209.39	8.05	0.65	0.6	0.52	0.654	0.808	0.692	CHY
CAV2 cluster	7	13	175.32	13.49	0.76	0.64	0.67	0.923	0.462	1	CY
DEFB1 cluster	8	20	178.86	8.94	0.26	0.37	0.19	0.9	0.85	0.9	CHY
LYN	8	24	130.76	5.45	0.66	0.44	0.49	0.75	0.333	0.875	CY
IFNB1cluster	9	48	387.84	8.08	0.58	0.55	0.54	0.479	0.667	0.875	HY
SYK	9	20	94.43	4.72	0.6	0.66	0.41	1	0.8	0.8	CHY
STOM	9	16	31.13	1.95	0.83	0.65	0.46	1	0.625	0.938	CHY
MBL2	10	6	6.32	1.05	0.74	0.66	0.28	0.667	0.667	1	CHY
THY1	11	5	2.76	0.55	0.48	0.61	0.38	1	0.8	1	CHY
CD4	12	10	31.26	3.13	0.34	0.38	0.16	1	1	0.9	CHY
CCNT1	12	6	24.03	4.00	0.71	0.83	0.62	0.333	0.167	1	Y
IFNG	12	7	40.83	5.83	0.96	0.86	0.72	0.571	0.571	0.857	
STOML1	15	8	9.14	1.14	0.54	0.53	0.53	0.625	0.625	0.75	CHY
ABCC1	16	30	192.77	6.43	0.51	0.55	0.3	0.933	0.9	0.933	CHY
IL4R cluster	16	14	182.83	13.06	0.45	0.33	0.15	0.857	0.571	0.929	CHY
CCL22 cluster	16	6	155.64	25.94	0.39	0.32	0.14	0.833	0.833	1	CHY
HP	16	4	6.43	1.61	0.29	0	0	1	0.5	0.25	CHY
NOS2A	17	13	43.77	3.37	0.53	0.57	0.54	1	1	1	CHY
FLOT2	17	5	18.21	3.64	0.69	0.71	0.49	0.8	0.8	0.8	CHY
CCL2 cluster	17	34	133.34	3.92	0.69	0.7	0.6	0.765	0.471	0.706	CY
MMP28 cluster	17	45	1029.74	22.88	0.65	0.65	0.48	0.689	0.689	0.844	CHY
PHB	17	10	10.82	1.08	0.53	0.74	0.26	0.4	1	0.8	HY
ICAM2	17	11	4.15	0.38	0.26	0.3	0.2	0.727	0.727	0.909	CHY
LMAN1	18	10	29.37	2.94	0.49	0.57	0.54	1	1	0.7	CHY
NFATC1	18	8	135.26	16.91	0.04	0.08	0.06	0.5	0.625	1	HY
VAV1	19	23	84.65	3.68	0.56	0.5	0.31	0.87	0.652	0.913	CHY
ICAM1cluster	19	58	642.32	11.07	0.51	0.5	0.49	0.638	0.69	0.914	CHY
TGFB1	19	9	23.56	2.62	0.54	0.64	0.51	0.667	0.667	0.889	CHY
FUT2	19	13	23.99	1.85	0.62	0.02	0.69	0.385	0.308	0.692	
RNPC2	20	14	38.47	2.75	0.67	0.65	0.77	0.214	0.071	0.643	
SLPI	20	4	2.33	0.58	0.55	0.69	0.25	0.75	1	1	CHY
IFNAR2 cluster	21	39	521.46	13.37	0.55	0.55	0.43	0.718	0.41	0.769	CY
APOBEC3G	22	8	10.61	1.33	0.7	0.69	0.64	0.75	0.375	1	CY

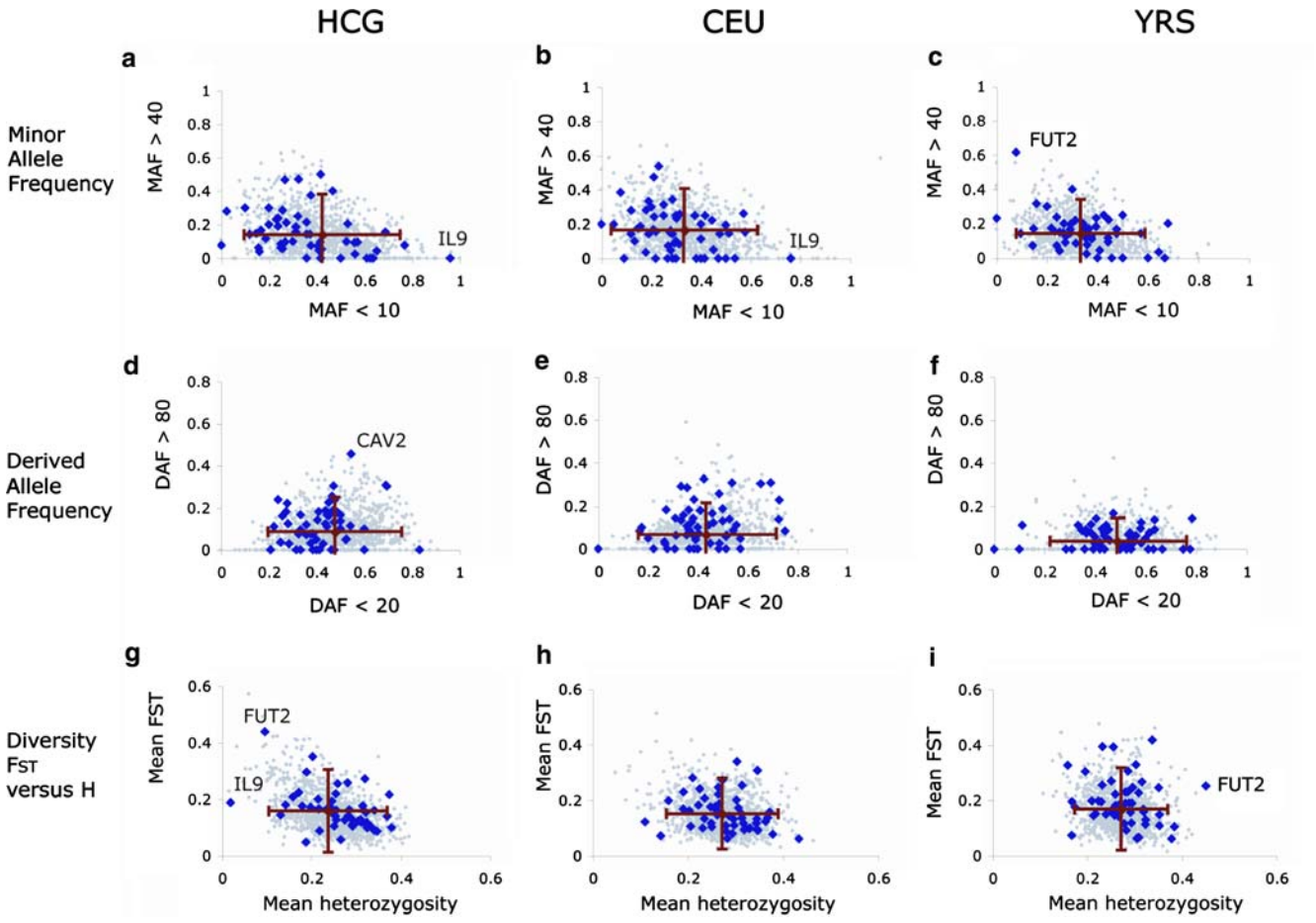

Fig. 1 Cross population analysis for signals of selection based on regional allele frequency differences. Panels **a**, **d**, and **g** represent data from the Chinese dataset, panels **b**, **e**, and **h** are from the European dataset, and panels **c**, **f**, and **i** are from the Yoruba dataset. In all panels *gray dots* indicated simulated data (Schaffner et al. 2005), *blue dots* indicate each of the 64 genes/regions under study, and *red lines* indicate average and 95% confidence intervals based on the simulated data. Panels **a**–**c** plot the percentage of SNPs with a MAF above 40% versus the percentage of SNPs with a MAF below 10%. IL-9 is an outlier in HCG (p-value = 0.029). FUT2 is an outlier in YRS (p-value = 4.09E-05). Panels **d**–**f** plot the percentage of SNPs with a DAF above 80% versus the percentage of SNPs with a DAF below 20%. CAV2 is an outlier in HCG (p-value = 2E-04) Panels **g**–**i** plot the average Fst for a region against the average heterozygosity for a region. FUT2 is an outlier in HCG (p-value = 0.003) and YRS (p-value = 0.008). IL-9 is an outlier in HCG (p-value = 0.025)

possible selective sweep in a single population, or (b) regions with high heterozygosity and high F_{ST} which may signal population-specific balancing selection.

The *relative extended haplotype homozygosity* (*REHH*) test compares, as a function of frequency, the length of haplotype identity around each allele as compared to an expected distribution (Fig. 2a–c). Under random drift, new variants require a long time to reach high frequency in the population, allowing the length of haplotype sharing time to decay due to recombination (Kimura 1983). Under positive selection, the time to becoming common is foreshortened, such that selected variants will carry unusually long haplotypes as compared to other variants of the same frequency. Relative EHH uses the extent of LD of other haplotypes at each locus as a control for local variation in recombination rate.

Empirical and expected distributions for each frequency test are shown in Fig. 1. The first observation is that the empirical and simulated data are largely overlapping, indicating that the population genetic simulations (which were parameterized using completely independent data) provide a good fit to the empirical results. This agreement then allows examination of the genes whose behavior appears to be an outlier relative (a) to the simulated data and (b) relative to the other members of the 168 genes examined.

Based on tests of allele frequency, three of the 168 genes stand out: IL-9, FUT2, and CAV2. At IL-9, the signal is due to a single common haplotype (AAATGA) that is very common in CEU and HCG (81.2 and 98.3%, respectively), but half as frequent in YRS (at 43.8%) (Figs. 1a, b, and g, Fig. 3a, and http://www.broad.mit.edu/personal/pardis/CandidatePaper/). The IL9 gene is an interleukin family member that has been shown to influence T-cell and mast cell proliferation (Demoulin and Renauld 1998) however little is known about its role in human disease.

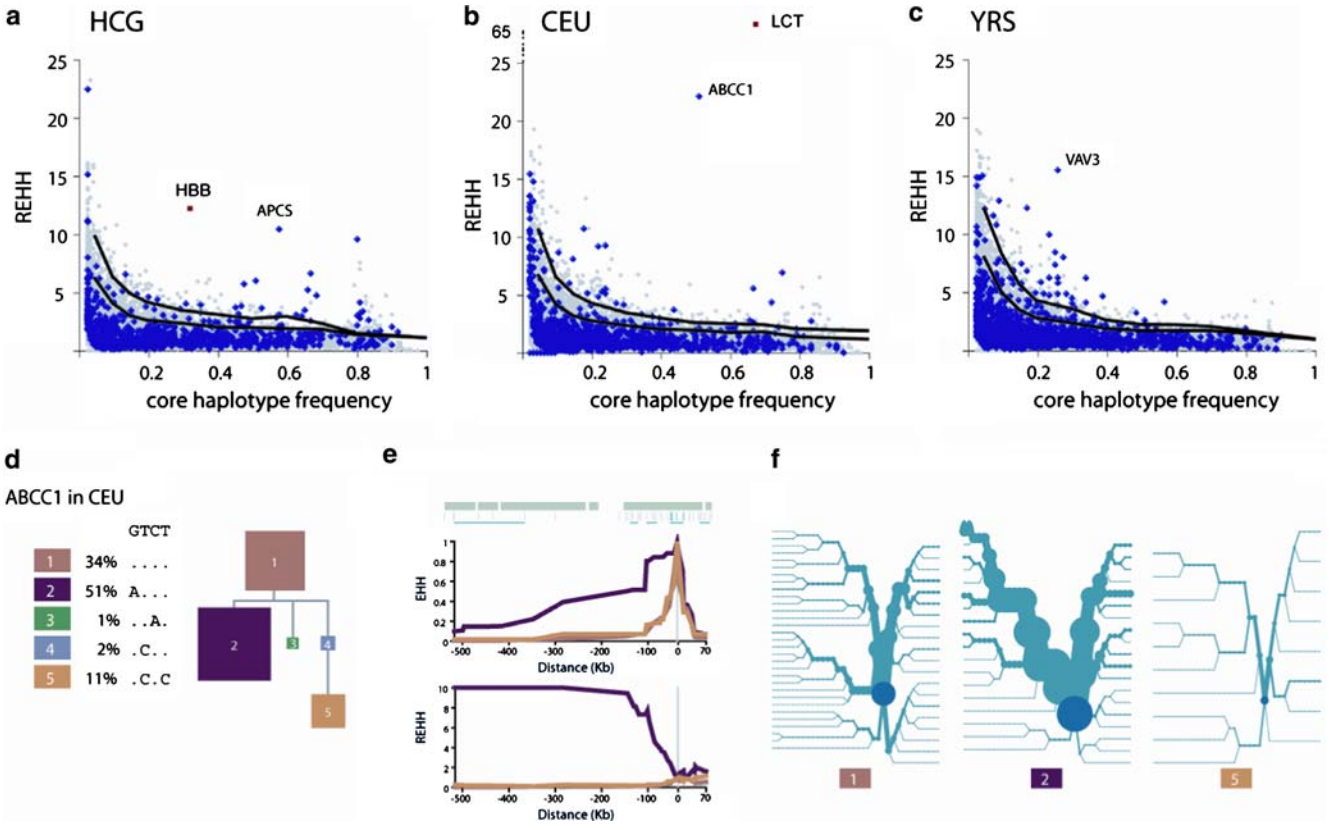

Fig. 2 REHH versus frequency distribution. In each population, we examined every variant of every haplotype block in the 168 gene dataset. Here we plot the REHH (REHH—a measure of haplotype specific LD, Materials and methods) against the haplotype variants frequency at 0.04 marker breakdown from the core. *Lines* represent from bottom to top the 95th and 99th percentiles. As reference we indicate positive controls from independent studies using HapMap data (*red squares*): HBB from HapMap YRS and LCT in HapMap CEU. **a**. A haplotype variant in APCS shows an increased REHH value given its frequency in the HCG (p-value = 1.72E-05). **b**. ABCC1 is an outlier in the CEU (p-value = 4.86E-10). **c**. A variant in VAV3 is an outlier in the YRS (p-value = 5.89E-07). **d-f**. Analysis of ABCC1 structure in the CEU. **d**. Frequencies and phylogenetic tree of haplotypes at ABCC1 (GTCT defined as ancestral haplotype based on consensus allele for three primates, size of *box* indicates frequency of haplotype). **e**. EHH and REHH plotted for each core haplotype at each distance from the core region. **f**. Haplotype birfurcation diagrams of the cores at ABCC1. The 51% maroon haplotype has apparently extended LD compared to the 34 and 11% haplotype at the locus. (EHH p-value = 0.044)

In FUT2, we observe a large number of high-frequency SNPs, and high heterozygosity, in the YRS sample (Fig. 2c), but low heterozygosity for this region in the HCG (with correspondingly high F_{ST} between YRS and HCG). Of 13 SNPs polymorphic in at least one population, only four are polymorphic in the HCG and only three of those are commonly polymorphic across the three populations (Fig. 3b). The most common three-marker haplotype in European chromosomes (GGC 58.8%) is absent in HCG and rarer in YRS (12.3%). Previous studies of FUT2 (Koda et al. 1996, 2003; Liu et al. 1999) have shown increased diversity between subpopulations, high F_{ST} between Asian and non-Asian populations, and elevated measures of heterozygosity in Europeans and Africans (by Tajima's D). The FUT2 gene encodes a protein responsible for determining an individuals' A, B, and O blood protein secretor status. Numerous association studies have been performed and there is some suggestion that variation in this gene may be protective against respiratory pathogens, helicobacter pylori, and HIV-1 (Ali et al. 2000; Hamajima 2003; Raza et al. 1991; Schaeffer et al. 2001).

The CAV2 cluster, containing the functionally related CAV1 and CAV2 genes, showed the strongest signal for selection in the DAF test in HCG (Fig. 1d). The haplotype structure of this region in the three populations reveals the nature of this outlier (Fig. 3c). Specifically the derived haplotype, TTTTTCT, is much more frequent in the HCG than the CEU or YRS, suggesting a possible selective sweep in Asia. The CAV2 and CAV1 genes encode Caveolin proteins which function as scaffolding structures in cholesterol rich lipid rafts in many cell types (neurons, immune cells, and endothelium) (Cohen et al. 2004).

The most striking result in the REHH test is a haplotype of ABCC1 that shows unusual extent in European samples (Fig. 2b, d). ABCC1 (or multi-drug resistance protein 1, MRP1) is one of a family of multi-specific organic anion transporters, which can protect cells against negative or neutrally charged drugs (Wijnholds 2002). Mouse models have suggested a role for this

a IL9

Haplotype alleles
for SNPs polymorphic in all 3 populations

CEU		HCG		YRS	
AAATGA	0.812	AAATGA	0.983	AAATGA	0.438
AACCCG	0.063	CGCCCG	0.017	AAATGG	0.134
CGCCCG	0.042			AGCCCG	0.117
CAATGA	0.021			AAACGA	0.112
AAACGA	0.021			AACCCG	0.099
AAACCG	0.021			CGCCCG	0.050
				CAACCG	0.025

b FUT2

Haplotype alleles
for SNPs polymorphic in all 3 populations

CEU		HCG		YRS	
GGC	0.588	GGT	0.538	AGC	0.263
AGT	0.129	AGT	0.262	AAC	0.202
AGC	0.110	AAT	0.091	GGT	0.140
AAT	0.087	AGC	0.054	GGC	0.123
GGT	0.064	GAT	0.037	AGT	0.121
GAT	0.021			AAT	0.114
				GAT	0.038

c CAV2

Haplotype alleles
for SNPs polymorphic in all 3 populations

CEU		HCG		YRS	
TTTTTCT	0.426	TTTTTCT	0.783	TTTTTCT	0.393
TTTGGCT	0.255	AATGGCT	0.173	TTTGGCT	0.270
TTCTTTC	0.189	TTCTTTC	0.033	TTCTTTC	0.261
AATGGCT	0.117	ATTGGCT	0.010	TTCTGCT	0.033
TTTTTTC	0.013			AATGGCT	0.016

Fig. 3 Comparison of Haplotypes of IL9, FUT2, and CAV2 regions between populations. **a** IL-9- haplotype alleles for the six SNPs polymorphic in all three populations. Notice that the most common haplotype in CEU and HCG chromosomes is half as frequent in YRS chromosomes. **b** FUT2—Of 13 SNPs polymorphic in at least one population, only four are polymorphic in the HCG and only three of those are commonly polymorphic across the three populations. These SNPs are in strong LD in CEU and YRS but not in HCG (see Supplemental LD plots). The most common three-marker haplotype in European chromosomes (GGC 58.8%) is absent in HCG and rarer in YRS (12.3%). **c**. CAV2—Haplotype allele frequencies reveal a seeming reduction in haplotype diversity in the Han Chinese. The second most common haplotype in Europeans and Yoruba (TTTGGCT 25.5/27.0%) is absent in the Han Chinese

gene in the resistance to Streptococcus pneumoniae (Schultz et al. 2001). APCS, the serum amyloid P component gene, and VAV3, a hemopoeitic cell specific guanine nucleotide exchange factor involved in actin rearrangement, show possible extended LD given their frequency in HCG and YRS, respectively (APCS Fig. 2a and Supplemental Fig. 1a–c; VAV3 Fig. 2c and Supplemental Fig. 2d–f). One important note is that the Duffy locus is included in this dataset and does not show an REHH signal in this dataset or in similar data from the HapMap (not shown). This suggests that REHH may be limited in its ability to detect certain selection events. We examine the power of the REHH test with a simulated dataset in Supplemental Table 3.

In summary, by applying four tests of selection on each of 168 genes in three populations, we identified six outliers that provided unusual patterns of genetic variation that could be consistent with signatures of past selection. As noted above, these signals do not represent definitive evidence for selection, but simply highlight candidate regions that might justify greater scrutiny. In particular, as SNP ascertainment bias and statistical fluctuation can result in false positive signals of selection, genotyping, and resequencing in additional cohorts will be required.

As an initial assessment of whether statistical fluctuations explained these outliers, we examined the available data collected by HapMap for these six genes. While this dataset has very similar issues of SNP ascertainment bias, as an independent set of SNPs and (for the HCG and YRS) DNA samples HapMap provides information with regard to statistical fluctuation.

Of the loci identified by tests of frequency distribution, none were strongly confirmed in the HapMap data. No signal was observed at CAV2 nor Il-9, suggesting either (a) that our original signal was a statistical fluctuation or due to ascertainment bias, or (b) that non-overlapping ascertainment issues in the HapMap data may have obscured a true signal. Perhaps not unexpectedly (given the previous literature on the gene) F_{ST} values were again moderately elevated at FUT2 in the HapMap data (0.25 CEU|HCJ, 0.23 CEU|YRI, 0.26 HCJ|YRI), although not as extremely as in the data we collected. These results are not particularly striking in comparison to the overall distribution of HapMap data (highest rank of the three was 708 of 9,704, empirical p-value = 0.07, for the CEU|HCJ).

In contrast, the outliers in the EHH/REHH test were again observed in the HapMap data. We observed a nine SNP haplotype at ABCC1 in Europeans that overlaps with the area previously highlighted, which shows a similar signal: a 44% haplotype with extended LD (EHH of 0.28, p-value = 0.024; REHH of 5.1, p-value = 0.007). While the individuals used in this study overlap largely with the HapMap European samples, the SNPs employed are distinct. The VAV3 extended haplotype signal in Yoruba is also observed in the HapMap Yoruba dataset. Specifically, we observe a nine SNP haplotype at VAV3 at with a statistically significantly extended haplotype similar to what we report here (frequency of 30%; EHH of 0.41, p-value = 0.006; REHH of 6.7, p-value = 0.002). For APCS we found an 11 SNP haplotype at 43% also significantly extended by REHH (EHH of 0.22, p-value = 0.085; REHH of 6.6, p-value of 0.015). These data demonstrate that all three EHH/REHH results are outliers when compared to the genome-wide empirical distribution from the HapMap project.

This characterization of 168 genes supports a number of conclusions, and suggests hypotheses for future research. First, for laboratories interested in association studies of genes involved in the immune response, our dataset offers additional data (along with HapMap and the Perlegen data) to select tag SNPs for association

studies. Second, we have piloted the search for signatures of evolutionary selection in HapMap-style data. Such searches are now possible genome-wide, and although fraught with difficulties because of ascertainment bias, nevertheless might hasten the discovery of medically important genes that were subject to evolutionary selection. In this regard, our experience may offer some guidance. Specifically, we found that the outliers in tests of allele frequency distribution were not confirmed in a second dataset performed under a similar design. This is consistent with the widespread understanding that there are many confounders in the ascertainment of SNPs in HapMap-style characterization, and that these ascertainment issues may mimic or obscure any signal of allele frequency distribution that may exist.

In contrast, the outliers in the test of haplotype length (REHH) were more robust in the second dataset. This may not be surprising as the assessment of haplotype length is most dependent on the ability to detect recombination events in the data, which we expect to be less sensitive to the affect of ascertainment bias on the allele frequency distribution (given adequate marker density).

We have provided data to suggest that six genes are more likely candidates to have undergone evolutionary selection, and thus could justify further analysis to confirm or deny whether this is the case. Interestingly, this could be taken to imply that most of the 168 immunological candidates have not undergone recent selective events. However another possibility is that our dataset and analyses are limited in their ability to identify cases of selection that are not extreme. We favor this interpretation and believe that one likely limitation is that SNP coverage in a subset of our regions may limit our power to detect a signal based on the frequency-based methods we use here. Moreover our power calculations in Supplemental Table 3 reveal limitations in the REHH method for certain selection scenarios full such as selective sweeps.

To further investigate our potential hits, as well as, the possible false negativity of most of the 168 genes, we turned another traditional test of selection: the comparison of rates of non-synonymous versus synonymous mutation in the coding regions (dN/dS ratios). This method is also limited by current SNP discovery but provides another measure of selective pressure based on hypotheses about functionality of coding SNPs.

A recent study of dN/dS ratio differences between chimpanzees and humans in over 20,000 genes has been published (Nielsen et al. 2005). In this study Nielsen et al. report increased non-synonymous mutations in immunity and defense related genes. Only 125 of our 168 selected genes were included in the Nielsen study and of that only 63 met inclusion criteria for analysis (>50 bp and >2 mutations reported). None of these 63 genes had a significant p-value, including three of the six possibly selected genes we discuss here: VAV3, APCS, and ABCC1. The CAV2 and FUT2 loci were not analyzed by Nielsen but were analyzed in an earlier study by Clark et al. (2003). Clark found no significant evidence at these two loci. The IL9 locus was not studied by Nielsen or Clark. This does not necessarily rule out a selective event at these genes. Rather this data may suggest either non-neutrally evolution in both chimpanzees and humans or, more likely, that the putative selected variation does not lie in the coding sequence of these genes. In either case further analysis will be required to confirm the possibility of recent evolutionary selection at these six genes, and in particular APCS, VAV3, ABCC1, and FUT2.

Materials and methods

Description of population samples

All YRS participants were recruited from a Yoruba-speaking rural community in Southwest Nigeria as part of a larger study of the genetics of hypertension. Study protocols were reviewed and approved by the institutional review boards at University College Hospital, Ibadan, and Loyola University Medical School. Written informed consent was obtained from the participants. The detailed sampling frame was described in our previous studies (Cooper et al. 1997, 2002). All CEU samples obtained from the Coriell Cell Repository and drawn from the collection of Utah CEPH pedigrees of European descent (see Coriell website, Electronic-Database Information). All the HCG samples were recruited from volunteers between April and June 2001. Two sites were chosen in Guangxi Province, southern China for a study examining the genetics of Nasal Pharyngeal Carcinoma. Study protocols were reviewed and approved by Chinese Human genetic resource committee and the National Cancer Institute. These volunteers were recruited from the Cancer Research and Control Institute in Wuzhou City and the Cangwu County Cancer hospital, located about 20 miles from Wuzhou. Written informed consent for the HCG samples was obtained from all study participants or a legal guardian for participants under the age of 18.

SNP Selection

The SNPs were selected from public databases in multiple batches over a 1.5 year period from July 2002 to December 2003 (indicated with "rs" names). Preference was given to "double hit" SNPs (85% qualify as double hit SNPs based on current database, 46% have been typed as part of the International Haplotype Map project). A small subset of SNPs was derived from Perlegen (12.7%) and Celera (2.5%) databases which were at the time proprietary and are now publicly available (indicated with "pr" and "hCV" names in our dataset).

Simulations

We used a computer program that simulates gene history with recombination based on a neutral model of

evolution described elsewhere (Schaffner et al. 2005; Hudson 1990). The published model includes a demographic model and a model of non-uniform recombination, and was designed to generate simulated data approximating empirical results for allele frequency distributions, population differentiation and linkage disequilibrium; its parameters were tuned for populations similar to those studied here (CEPH, Yoruba, and a mixture of Chinese and Japanese). For the present study, we simulated 1 Mb segments of DNA, one for each sample in the three populations. The SNP ascertainment was modelled on the selection strategy used by the SNP Consortium (Sachidanandam et al. 2001), which was the source of most of the SNPs in our study. For this purpose, a subset of the simulated chromosomes was withheld from analysis and used to create a simulated reference genome and low-coverage shotgun reads; sites that differed between reference and shotgun read were treated as having been ascertained. We did not attempt to model the bias toward double-hit SNPs. This bias has little effect on the FST and derived frequency tests, and means that the tests for heterozygosity and excess low-frequency alleles are conservative; only for the excess high-frequency alleles test is the simulation anticonservative.

We generated 1,000 simulated loci, each containing a full set of chromosomes for each population. Within these loci we identified regions 120 kb in length with one SNP every 4 kb to compare to the genic regions. At longer distances, we chose SNPs one every 40 kb from the core region for up to 400 kb.

Sequenom

The SNP primers and probes were designed in multiplex format (average fivefold multiplexing) with SpectroDESIGNER software (Sequenom), as previously described (Walsh et al. 2003).

Illumina

Genotyping for this project was performed at Illumina Inc., using the BeadLab system. The SNPs were selected and screened with Illumina's bioinformatics software, which is used to predict optimal oligonucleotide probe sequences for each marker. Each potential set of SNP assay oligonucleotides was evaluated for characteristics that included G-C content, melting temperature, self-complementarity, and uniqueness in the genome. Based on the results of this evaluation 1,172 designable SNP markers (in an attempt at covering 576 individual SNP loci) were selected for assay design and validation testing. These loci were multiplexed in a single reaction using Illumina's GoldenGate® assay which has been previously described by Fan et al (2004). Out of these designable SNPs, data was returned for 547 out of the requested 576 loci.

ParAllele

Once the batch of SNPs in and around the candidate genes was selected, a homology sequence was selected based on *Tm* optimization. This optimization produced homology sequences around 40.4 bases long centered over the SNP and complementary to the genomic sequence surrounding it. This homology sequence was BLASTed against the genome to determine whether the sequence was unique in the genome (unique is defined as an exact match to only one position). If the exact sequence appeared more than once, the probe was not synthesized. This was the only filter used. Next, a tag sequence unique among the batch of assays and complementary to a feature on the detection chip system was added. No consideration was given to the degree of complementarity of sequences within the batch. Thus, probes can target overlapping sequences since the genomic DNA is not saturated with hybridized probe. All probe batches were manufactured by ParAllele Bioscience (South San Francisco, CA, USA) using its proprietary MIP probe synthesis procedures and are commercially available (MegAllele™ kit, ParAllele BioScience). This process is a pooled procedure, which results in a pool of up to 12,000 probes, which are tested using pooled QC procedures. The genotyping reactions were carried out using the standard protocols recommended by the manufacturer at the NCI LGD. The MIP assays are carried out in 96 well plates using 12 individuals per plate for each of four allele channels using the MegAllele genotyping kit (ParAllele Bioscience). Each of the four MIP reactions was mixed with 500 ng of genomic DNA. These SNPs were assayed using a four color MegAllele assay kit from ParAllele Bioscience. The scanner used is a CCD camera with a broadband excitation lamp (GeneChip AT Scanner, Molecular Devices, Union City, CA, USA). Data from each chip is collected using four different filters to yield the four images that are processed in a manner identical to the data from the two chips that are used in the two color protocol.

Snap

Snap is a Java-based program, derived from Haploview (Barrett et al. 2005), that allows for the visualization and analysis of genotype data from multiple populations and multiple regions of the genome simultaneously (Fry et al in preparation). It displays genes and SNPs in each region, SNP genotyping success, D' plots (Lewontin 1995), haplotype blocks, block tagging SNPs, the correlation coefficient r^2 for all regions and all populations. Using Snap we created D' plots and haplotype variation plots for all genes and populations.

Minor allele frequency analysis

We calculated the MAF for all SNPs genotyped successfully in all three populations. We broke up gene

clusters into regions of up to 40 SNPs and 160 kb. All allele frequency analysis was carried out on 1,000 independent simulated regions of roughly 30 SNPs over 120 kb. We assessed the fraction of SNPs for each region that have MAF less than 10% and the fraction of SNPs that have MAF greater than 40%.

Derived allele frequency

We calculated the DAF for all SNPs where ancestral allele status could be determined with high likelihood by genotyping a representative chimpanzee, gorilla, and orangutan. If there were a consensus primate allele across all successfully genotyped primates, it was identified as the ancestral. Otherwise, no ancestral allele was determined. We assessed the fraction of SNPs for each region that have DAF less than 20% and the fraction of SNPs that have DAF greater than 80%.

F_{ST} and H

Mean pairwise distance fixation index F_{ST} was used to calculate genetic differentiation between the three populations (Akey et al. 2002; Nei and Chesser 1983). F_{ST} partitions the total variance into within and between population components, quantifying the inbreeding effect of population substructure.

Nei's measure of heterozygosity (Nei 1987) the probability that any two randomly chosen samples from a population are the same, was used to calculate SNP diversity:

$$\pi = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_{2\ i}\right)$$

where n is the number of gene copies in the sample, k s the number of haplotypes, and p_i is the frequency of the ith haplotype.

Allelle frequency p-values

p-values were calculating using the mean and standard deviation of the simulated regions. We did a Bonferroni correction for the 55 regions tested with adequate coverage across the local gene region.

EHH and REHH analysis

Extended haplotype homozygosity (EHH) assesses the age of each haplotype at a gene by measuring the decay of the extended (SNPs far away from the gene) ancestral haplotype, which occurs over time with recombination. For a population of individuals sharing core haplotype X, 'EHH' is the probability that any two randomly chosen samples of core haplotype X have the same extended haplotype (Sabeti et al. 2002). The EHH is calculated as:

$$\mathrm{EHHH}_t = \frac{\sum_{i=1}^{s}\binom{e_{it}}{2}}{\binom{c_t}{2}}$$

where EHH is the extended haplotype homozygosity, t is the core haplotype tested, c is the number of samples of a particular core haplotype, e is the number of samples for a particular extended haplotype, and s is the number of unique extended haplotypes.

To correct for local variation in recombination rates, we compare the EHH of a tested core haplotype to that of other core haplotypes present at the locus, using the measure relative EHH (REHH). The REHH is the factor by which EHH decays on the tested core haplotype compared to the decay of EHH on all other core haplotypes combined. One must first calculate the '$\overline{\mathrm{EHH}}$', the decay of EHH on all other core haplotypes combined. For this we use the following equation where n is the number of different core haplotypes:

$$\overline{\mathrm{EHH}} = \frac{\sum_{j=1,j\neq 1}^{n}\left[\sum_{i=1}^{s}\binom{e_i}{2}\right]}{\sum_{i=1,i\neq 1}^{s}\binom{c_i}{2}}$$

The relative EHH (REHH) is simply $\mathrm{EHH}_t/\overline{\mathrm{EHH}}$.

EHH and REHH were calculated for all haplotypes in all haplotype blocks in the 64 studied regions and 1,000 simulated regions. Haplotypes were placed into 20 bins based on their frequency. p-values were obtained by log-transforming the EHH and REHH in the bin to achieve normality, and calculating the mean and standard deviation. We did a Bonferroni correction for the 64 regions tested. All analysis was carried out using the Sweep™ software program (PCS et al. in preparation).

REHH test power estimation

To estimate the power of the REHH method, we first simulated 500 regions of DNA under neutral evolution, 1 Mb in length, in data samples matching the 120 chromosomes in CEU, YRI, and HCB populations (Schaffner et al. 2005). We then modified the program to generate simulations of selective sweeps for 200 regions of DNA, 1 Mb in length, in 120 chromosomes in each of the three populations. We tested the power to detect selection for a variety of parameters including population examined, date of origin of the selected mutation, frequency attained by the selected mutation, and the genetic distance at which REHH is examined. Overall, our power to detect selection is greatest with the Yoruba and weakest with the Asian population, which is likely due to the greater degree of population bottlenecks in the Asian population. The REHH method, in its current

form also has greater power at detecting partial sweeps than complete sweeps, and is being modified to better detect complete sweeps (PV, BF, ESL, PCS unpublished data). The Supplemental Table 3 presents the fraction of simulations for which a signal of selection was identified at a genetic distance of 0.04 'observed historical recombination' using a strict cut off (p-value 0.0000316) for a variety of parameters of positive selection.

Acknowledgments E.C.W was supported by a Cancer Research Institute fellowship. PCS is funded by the Damon Runyon Cancer Research Foundation and by a L'Oreal Women in Science Award. H.B.H. is an NCI Cancer Research Training Award (CRTA) Postdoctoral Fellow. This work was funded through a special grant from the National Institutes of Health National Institute of Allergy and Infectious Disease. This publication has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract no. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was support in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814

Ali S, Niang MA, N'Doye I, Critchlow CW, Hawes SE, Hill AV, Kiviat NB (2000) Secretor polymorphism and human immunodeficiency virus infection in Senegalese women. J Infect Dis 181:737–739

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963

Cohen AW, Hnasko R, Schubert W, Lisanti MP (2004) Role of caveolae and caveolins in health and disease. Physiol Rev 84:1341–1379

Cooper R, Rotimi C, Ataman S, McGee D, Osotimehin B, Kadiri S, Muna W, Kingue S, Fraser H, Forrester T, Bennett F, Wilks R (1997) The prevalence of hypertension in seven populations of west African origin. Am J Public Health 87:160–168

Cooper RS, Luke A, Zhu X, Kan D, Adeyemo A, Rotimi C, Bouzekri N, Ward R (2002) Genome scan among Nigerians linking blood pressure to chromosomes 2, 3, and 19. Hypertension 40:629–633

Demoulin JB, Renauld JC (1998) Interleukin 9 and its receptor: an overview of structure and function. Int Rev Immunol 16:345–364

Fan JB, Yeakley JM, Bibikova M, Chudin E, Wickham E, Chen J, Doucet D, Rigault P, Zhang B, Shen R, McBride C, Li HR, Fu XD, Oliphant A, Barker DL,Chee MS (2004) A versatile assay for high-throughput gene expression profiling on universal array matrices. Genome Res 14(5):878–885

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

Hamajima N (2003) Persistent helicobacter pylori infection and genetic polymorphisms of the host. Nagoya J Med Sci 66:103–117

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70:369–383

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Hudson RR (1990) Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology, vol 7. pp1–44

Kimura M, (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge New York

Koda Y, Ishida T, Tachida H, Wang B, Pang H, Soejima M, Soemantri A, Kimura H (2003) DNA sequence variation of the human ABO-secretor locus (FUT2) in New Guinean populations: possible early human migration from Africa. Hum Genet 113:534–541

Koda Y, Soejima M, Liu Y, Kimura H (1996) Molecular basis for secretor type alpha(1,2)-fucosyltransferase gene deficiency in a Japanese population: a fusion gene generated by unequal crossover responsible for the enzyme deficiency. Am J Hum Genet 59:343–350

Lewontin RC (1995) The detection of linkage disequilibrium in molecular sequence data. Genetics 140:377–388

Liu YH, Koda Y, Soejima M, Pang H, Wang BJ, Kim DS, Oh HB, Kimura H (1999) The fusion gene at the ABO-secretor locus (FUT2): absence in Chinese populations. J Hum Genet 44:181–184

Nei M, (1987) Molecular evolutionary genetics (Eqn 8.4). Columbia University Press, New York

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. Ann Hum Genet 47(Pt 3):253–259

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J JS, Adams MD, Cargill M (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:e170

O'Brien SJ, Nelson GW (2004) Human genes that limit AIDS. Nat Genet 36:565–574

Raza MW, Blackwell CC, Molyneaux P, James VS, Ogilvie MM, Inglis JM, Weir DM (1991) Association between secretor status and respiratory viral illness. BMJ 303:815–818

Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. Nat Genet 33:457–458

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

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Schaeffer AJ, Rajan N, Cao Q, Anderson BE, Pruden DL, Sensibar J, Duncan JL (2001) Host pathogenesis in urinary tract infections. Int J Antimicrob Agents 17:245–251

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation. Genome Res 15(11):1576–1583

Schultz MJ, Wijnholds J, Peppelenbosch MP, Vervoordeldonk MJ, Speelman P, van Deventer SJ, Borst P, van der Poll T (2001) Mice lacking the multidrug resistance protein 1 are resistant to Streptococcus pneumoniae-induced pneumonia. J Immunol 166:4059–4064

Taylor MF, Shen Y, Kreitman ME (1995) A population genetic test of selection at the molecular level. Science 270:1497–1499

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Walsh EC, Mather KA, Schaffner SF, Farwell L, Daly MJ, Patterson N, Cullen M, Carrington M, Bugawan TL, Erlich H, Campbell J, Barrett J, Miller K, Thomson G, Lander ES, Rioux JD (2003) An integrated haplotype map of the human major histocompatibility complex. Am J Hum Genet 73:580–590

Wijnholds J (2002) Drug resistance caused by multidrug resistance-associated proteins. Novartis Found Symp 243:69–79; discussion 80–82, 180–185