A Genome Scan to Detect Candidate Regions Influenced by Local Natural Selection in Human Populations

Manfred Kayser, Silke Brauer, and Mark Stoneking

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

As human populations dispersed throughout the world, they were subjected to new selective forces, which must have led to local adaptation via natural selection and hence altered patterns of genetic variation. Yet, there are very few examples known in which such local selection has clearly influenced human genetic variation. A potential approach for detecting local selection is to screen random loci across the genome; those loci that exhibit unusually large genetic distances between human populations are then potential markers of genomic regions under local selection. We investigated this approach by genotyping 332 short tandem repeat (STR) loci in Africans and Europeans and calculating the genetic differentiation for each locus. Patterns of genetic diversity at these loci were consistent with greater variation in Africa and with local selection operating on populations as they moved out of Africa. For 11 loci exhibiting the largest genetic differences, we genotyped an additional STR locus located nearby; the genetic distances for these nearby loci were significantly larger than average. These genomic regions therefore reproducibly exhibit larger genetic distances between populations than the ''average'' genomic region, consistent with local selection. Our results demonstrate that genome scans are a promising means of identifying candidate regions that have been subjected to local selection.

Introduction

Analyses of genetic and morphological variation in human populations have demonstrated that a major event in human evolution was a recent African origin of modern humans, followed by dispersal out of Africa (Stoneking 1993; Lahr and Foley 1998; Mountain 1998). As human populations spread around the globe, they encountered a variety of novel selective pressures (e.g., new environments, climates, diets, parasites, and diseases), which must have required some genetic response. Yet, selection has been definitely shown to influence variation at only a few genes in human populations, almost all of which involve resistance to malaria (Allison 1954; Luzzatto, Usanga, and Reddy 1969; Miller et al. 1976; Hill et al. 1991; Tishkoff et al. 2001; Hamblin, Thompson, and Di Rienzo 2002). Identifying additional genes that have been subjected to different selection pressures in different populations (i.e., local selection) would greatly enhance our knowledge of the ways in which genetic variation in human populations has been shaped by natural selection, but detecting such genes is a challenge (Schlötterer 2002*b*).

One approach, suggested many years ago, is based on the idea that genes that are subjected to local selection should exhibit larger than average genetic distances between populations (Cavalli-Sforza 1966). Since this should also be true for marker loci closely linked to the selected locus, screening random marker loci across the genome for large genetic distances might be a useful way to identify genomic regions under local selection. Indeed, Lewontin and Krakauer (1973) proposed a statistical test, based on the expected variance in F_{st} values for a sample of loci, to detect loci with significantly large F_{st} values. Unfortunately the test was flawed (Lewontin and Krakauer 1975; Nei and Maruyama 1975; Robertson 1975), and the approach was largely abandoned. However, although the specific test is not valid, the general idea may still have merit. In fact, other authors have suggested that selection may be responsible for particular observations of loci exhibiting large genetic distances between populations (Bowcock et al. 1991), and recently there has been a resurgence of interest in methods to detect such loci (Beaumont and Nichols 1996; Baer 1999; Vitalis, Dawson, and Boursot 2001; Akey et al. 2002; Balloux and Goudet 2002; Payseur, Cutter, and Nachman 2002; Schlötterer 2002a). In this paper we investigate the feasibility of a genome scan approach to detect marker loci that exhibit large genetic distances between human populations, as a means of identifying candidate genes that have experienced local selection.

Key words: selection, humans, genetic distance, STR.

E-mail: stoneking@eva.mpg.de.

Mol. Biol. Evol. 20(6):893–900. 2003
DOI: 10.1093/molbev/msg092

Materials and Methods

Blood samples were obtained from 48 Europeans (blood donors from Leipzig, Germany) and 23 Africans (from Gondar, Ethiopia) and DNA was extracted by standard phenol/chloroform or salting-out procedures. An additional 24 African DNA samples (from the Nguni, Sotho-Tswanga, and Tsonga groups of South Africa) were provided by H. Soodyall. All samples were from unrelated individuals and were obtained with informed consent.

Genotyping was carried out at the Swedish Genome Center, Uppsala, using loci and methods as described previously (Lindqvist et al. 1996), with the following exceptions: amplifications were carried out in a 9 μl volume; PCR and pooling and dilutions of PCR products were performed with an ABI877 Integrated Thermal Cycler (PE Applied Biosystems, Inc.); and subsequent fragment length analysis was carried out with an ABI PRISM 3700 DNA Analyzer and GeneScan software (PE Applied Biosystems, Inc.). Genotypes are available from the authors upon request.

Additional candidate STR loci in particular genomic regions of interest were identified by screening the DNA sequence with the UCSC Genome Browser Gateway (Kent et al. 2002; http://genome.ucsc.edu/cgi-bin/hgGateway) for 14 or more copies of a dinucleotide and six or more copies of a trinucleotide or tetranucleotide repeated sequence. Primers were designed with the Primer3

Table 1
Average Heterozygosity and Number of Alleles by Population and Repeat Type for 332 STR Loci

		Africans		Europeans		Total	
STR Repeat Type	Number of Loci	Heterozygosity	Alleles	Heterozygosity	Alleles	Heterozygosity	Alleles
Dinucleotide	54	0.844	11.9	0.796	9.1	0.834	12.7
Trinucleotide	41	0.785	8.2	0.724	6.8	0.770	8.5
Tetranucleotide	237	0.782	8.0	0.756	7.1	0.776	8.5
TOTAL	332	0.792	8.6	0.759	7.4	0.785	9.2
$F_{2,329}$[a]		19.98***	60.79***	12.64***	21.25***	21.02***	64.99***

[a] F-value from one-way ANOVA with effect df = 2 and error df = 329.
*** $P < 0.001$.

program (http://www-genome.wi.mit.edu/genome_software/other/primer3.html), and fluorescent-labeled PCR was performed using standard conditions. The PCR products from up to three nonoverlapping STR loci were pooled and analyzed on an ABI 377 DNA Sequencer and GeneScan software. All loci have been submitted to the Human Genome Database (http://www.gdb.org), from which additional marker and typing details can be obtained. For some loci, direct DNA sequence analysis was performed using the Big Dye Reaction terminator Cycle Sequencing Kit and an ABI 377 DNA Sequencer (PE Applied Biosystems, Inc.).

The average heterozygosity, number of alleles per locus, tests for goodness of fit to Hardy-Weinberg proportions, and R_{st} values were calculated with the software FSTAT2.9.3 (http://www.unil.ch/izea/softwares/fstat.html). R_{st} is analogous to F_{st} but is based on a stepwise mutation model (Slatkin 1995), which is generally considered to be more appropriate for STR loci. Large R_{st} values indicate large genetic distances between populations. We also calculated the ln RV value for each locus, which is the natural log of the ratio of the variance in allele size for two populations (Schlötterer 2002*a*). Large positive or negative values of ln RV for a particular STR locus indicate that one population has a much smaller allele size variance, which in turn might reflect a recent selective sweep at a nearby locus in that population. One-way ANOVA and nonparametric tests were carried out with STATISTICA (Statsoft, Inc.).

Results

Genetic Diversity Within Populations

Usable results were obtained for 332 of the 351 loci genotyped; the remaining 19 loci either amplified weakly, gave nonspecific products, or gave fragment sizes inconsistent with expectations. The number of autosomal loci showing a significant excess or deficit of heterozygotes, relative to Hardy-Weinberg expectations, was within expectations for the German sample. However, among the Africans, 28 loci showed a significant deficit of heterozygotes, which is significantly more than the 15.8 loci expected ($P < 0.05$). This most likely reflects pooling of the Ethiopian and South African samples, as analyzing them separately eliminates the excess of loci exhibiting a significant deficit of heterozygotes. In all subsequent analyses we therefore treated the two African samples separately as well as together, and in no case were the conclusions altered by treating the samples separately.

The 332 loci included 54 dinucleotide repeats, 41 trinucleotide repeats, and 237 tetranucleotide repeats; both average heterozygosity and the number of alleles for both Africans and Europeans differed significantly with respect to repeat type (table 1). All measures of variation were significantly higher for dinucleotide repeats, based on one-way ANOVA. The Africans had significantly higher average heterozygosities and number of alleles per locus than the Europeans (Wilcoxon matched pairs test: heterozygosity, $Z = 9.44$, $P < 0.001$; number of alleles, $Z = 10.11$, $P < 0.001$). The difference remained significant when the Ethiopians and South Africans were analyzed separately (results not shown).

We also analyzed the average heterozygosity and number of alleles per locus for each chromosome (fig. 1). Based on one-way ANOVA, neither measure of variation differed significantly among chromosomes for either the Africans and Europeans separately or together, nor did they differ when the average of the autosomal loci was compared with the average of the X-linked loci (results not shown).

Genetic Differences Between Populations

The distribution of R_{st} values is shown in figure 2. The average R_{st} value was 0.043 and did not differ significantly with respect to repeat type ($F_{2,329} = 2.87$, $P > 0.05$). Although the average R_{st} value was less for autosomal loci (0.042) than for X-linked loci (0.064), the difference was not statistically significant (Mann-Whitney U test, $Z = -1.16$, $P > 0.2$).

The distribution of ln RV values is shown in figure 3. The average ln RV value was 0.20, which is significantly greater than 0 ($P < 0.001$), the value expected if the variance in allele size distribution is equal for Africans and Europeans. Since the variance in Africans appears in the numerator of the ln RV values, this is another indication of significantly greater genetic variation in Africans than in Europeans at these loci. As with the R_{st} values, the ln RV values did not differ significantly with respect to repeat type ($F_{2,329} = 0.56$, $P > 0.5$), and while the average ln RV value was less for autosomal loci (0.18) than for X-linked loci (0.50), the difference was not statistically significant ($Z = -1.55$, $P > 0.1$).

Searching for Genomic Regions Influenced by Local Selection

It would seem that the obvious way to identify loci exhibiting significantly large genetic distances would be to

compare the observed distribution of R_{st} and ln RV values with that expected under neutrality. However, the expected distribution of R_{st} values under neutrality is not known, and although simulations show that ln RV values tend to follow a normal distribution under a wide variety of demographic scenarios (Schlötterer 2002*a*), there is an additional complicating factor: the simulations assumed two independent (unrelated) populations with no gene flow, which is not true for any empirical situation, including the African and European populations compared here. Absence of independence is a crucial factor that precludes any simple comparison of the observed distributions to a neutral model that assumes such independence (Nei and Chakravarti 1977; Nei, Chakravarti, and Tateno 1977). We therefore adopted a strictly empirical approach and assumed that those loci with the highest R_{st} and/or ln RV values would be most likely to mark genomic regions that have been subjected to local selection. To identify such loci, we plotted the R_{st} versus ln RV values for each locus (fig. 4). The individual R_{st} and ln RV values are significantly correlated ($r = 0.27$, $P < 0.001$).

A complicating factor is that high R_{st} and/or ln RV values could result from the mutational process or chance events involving that specific STR locus. For example, one locus exhibits a strikingly low ln RV value of −1.82 (fig. 4), indicating a much greater allele size variance in Europeans than in Africans at this locus. Inspection of the allele size distribution for this locus (D20S173) immediately reveals the reason for this low ln RV value: a 134-bp allele, which is approximately 40 bp smaller than the next smallest allele, is found at a frequency of 15% in the Europeans but only 1% in the Africans (fig. 5). Consequently, this large size difference greatly inflates the variance in allele size in Europeans. We sequenced this allele and several others and found that the 134-bp allele reflects a single large deletion rather than a difference in the number of repeats at this locus (data not shown). Hence, the unusual ln RV value for D20S173 is caused by a unique mutation at this STR locus that does not conform to the stepwise mutation model.

One way to distinguish between such chance mutational events and local selection is to examine R_{st} and ln RV values at closely linked loci. If local selection on the genomic region is responsible for the high R_{st} and/or ln RV value, then other STR loci in the same genomic region should also exhibit high values. Conversely, if chance events at the STR locus are responsible for the high value, then we would not expect high values at closely linked loci. We therefore selected 15 loci that exhibited

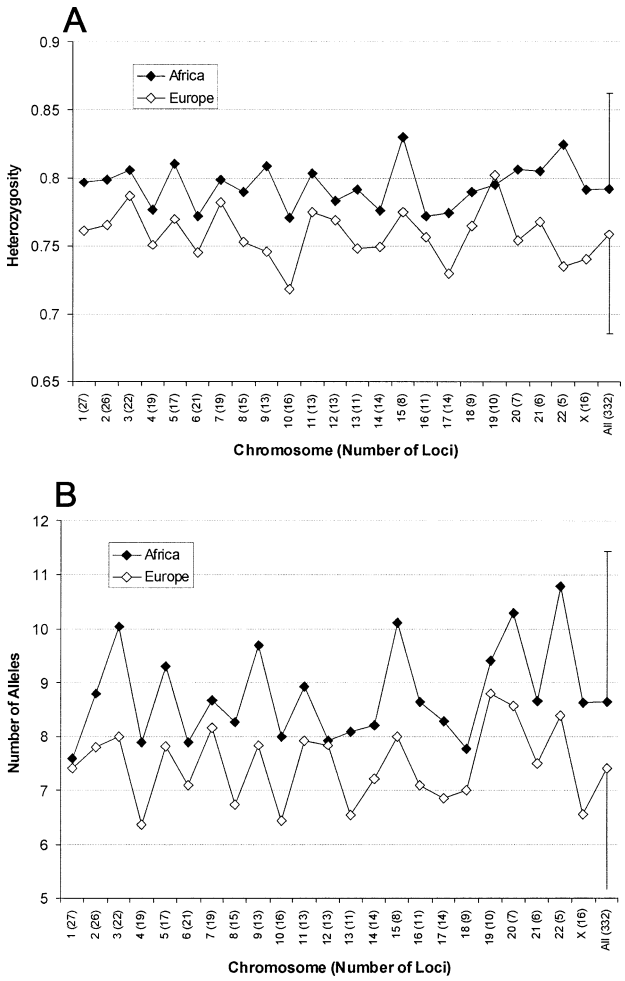

FIG. 1.—Average heterozygosity (*A*) and number of alleles (*B*) by population and chromosome.

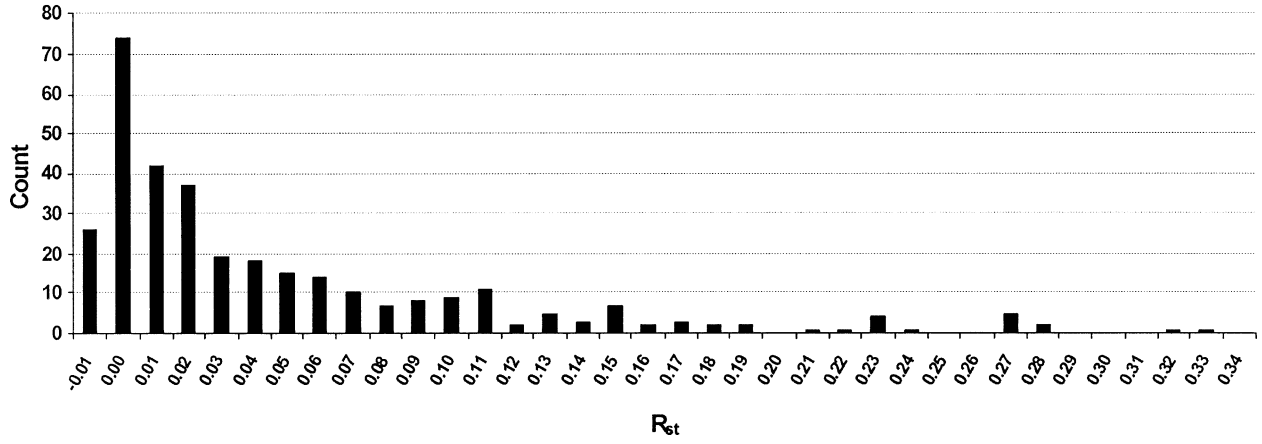

FIG. 2.—Distribution of R_{st} values between Africans and Europeans for 332 STR loci.

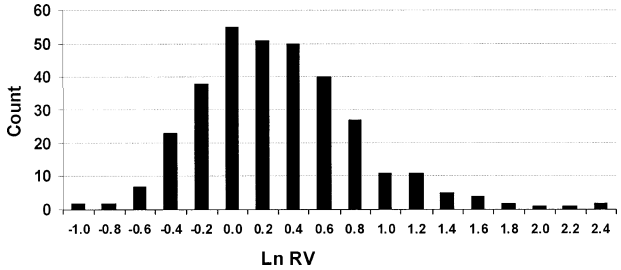

FIG. 3.—Distribution of ln RV values between Africans and Europeans for 332 STR loci.

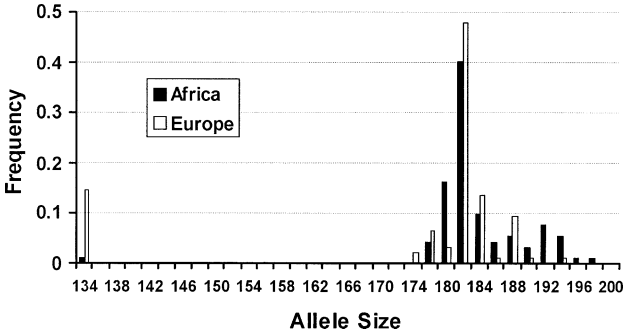

FIG. 5.—Allele size distribution for D20S173, a locus with an unusually low ln RV value (see fig. 4).

high R_{st} and/or ln RV values (fig. 4), located them in the human genome with the UCSC Genome Browser Gateway (Kent et al. 2002), and screened the genomic sequence within 50 kb of each of these 15 loci for additional candidate STR loci. For 11 of these ''target'' loci, we could identify a potential ''nearby'' STR locus; we then designed primers, genotyped the Africans and Europeans, and calculated R_{st} and ln RV values for these nearby loci (table 2). The average R_{st} value for these nearby loci was 0.171, and the average ln RV value was 1.15; both values are significantly greater than the average R_{st} and ln RV values for the original 332 loci (Mann-Whitney U test: R_{st}, $Z = -3.21$, $P < 0.01$; ln RV, $Z = -2.39$, $P < 0.05$). Thus, loci in the same genomic region as the target loci also exhibit unusually high R_{st} and/or ln RV values.

To further investigate the properties of these unusual genomic regions, we took one such region and searched for additional STR loci. The target locus, D2S1400, has both a high R_{st} value (0.317) and a high ln RV value (2.30), and maps 0.9 kb from the gene for the E2F transcription factor 6 (E2F6; RefSeq ID NM 001952). An additional four STR loci were characterized in this region (fig. 6); R_{st} and ln RV values are high only for an additional locus (D2S3021) in the immediate vicinity of the E2F6 gene, suggesting that this gene might have been a target for local selection.

Discussion

We genotyped 332 STR loci, spaced at roughly 10-cM intervals, in a sample of Africans and a sample of Europeans. Dinucleotide repeat loci exhibited significantly higher variability than did loci with either trinucleotide or tetranucleotide repeats, in accordance with the view that dinucleotide repeat loci have a higher mutation rate (Chakraborty et al. 1997). We also found significantly higher levels of variability in the African sample, consistent with many previous genetic studies of human populations (Vigilant et al. 1991; Stoneking et al. 1997; Jorde et al. 2000; Yu et al. 2002), indicating a longer history and/or larger size for African populations.

There were no significant differences between autosomal loci and X-linked loci for variability within populations. This contrasts with estimates of nucleotide diversity based on DNA sequence analysis, in which lower levels of variability were found for X-linked loci (Yu et al. 2002), although the statistical significance of the difference in variability was not reported. Similarly, another study (Payseur, Cutter, and Nachman 2002) recently found weak (i.e., statistically nonsignificant) evidence of lower diversity on the X chromosome for published data on STR loci in Europeans. The smaller effective population size of the X chromosome relative to the autosomes leads to the expectation that diversity should be lower for X-linked loci, although the extent of this reduction also depends on the extent to which there is differential reproduction among males (Caballero 1995); the fewer the number of males that reproduce each generation, the more equal the effective sizes for X-linked and autosomal loci. Our failure to detect differences in variability for X-linked versus autosomal loci may reflect the high intrinsic mutation rate of STR loci, or it may reflect other factors such as differential reproduction among males.

Similarly, we found no significant differences between autosomal and X-linked loci with respect to genetic differentiation between populations, although both R_{st} and ln RV values were bigger for the X-linked loci than for the autosomal loci. However, three of the 15 loci that we identified for further study because of high R_{st} and/or ln RV values are on the X chromosome, significantly more than expected by chance ($\chi^2 = 4.81$, df = 1, $P < 0.05$). Payseur, Cutter, and Nachman (2002) also found evidence for more departures from neutrality for STR loci on the X chromosome. They interpreted this as evidence for more positive selection on the X chromosome, but since the data they analyzed came from a single population source (Europeans), their results could reflect either local selection (i.e., affecting only Europeans) or positive selection (i.e., affecting all human populations). Larger

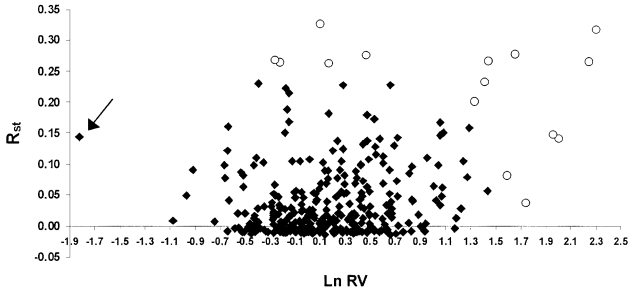

FIG. 4.—Plot of R_{st} versus ln RV values for 332 STR loci. The arrow points to the values for D20S173, a locus exhibiting an unusually low ln RV value. The open circles indicate loci with high R_{st} and/or ln RV values that were selected for further examination.

Table 2
Characteristics of Target Loci with Large R_{st} and/or ln RV Values and of Nearby Loci

Original Locus	H	A	R_{st}	ln RV	Nearby Locus	H	A	R_{st}	ln RV	Distance Away (kb)
D13S173	0.888	15	0.233	1.409	D13S1852	0.609	9	0.047	1.582	11.2
D15S657	0.728	10	0.037	1.739	D15S1541	0.738	8	−0.001	0.791	3.4
D16S539	0.804	8	0.268	−0.265	D16S3418	0.786	8	0.422	−0.669	13.7
D7S550	0.852	14	0.202	1.329	D7S3249	0.733	12	0.307	1.357	12.2
D21S1437	0.818	10	0.263	0.170	D21S2095	0.704	8	0.079	−0.272	0.2
D9S2169	0.745	7	0.266	1.440	D9S2182	0.845	13	0.202	1.915	27.3
D2S1400	0.757	12	0.317	2.299	D2S3021	0.746	7	0.378	0.536	12.5
D6S1031	0.845	9	0.277	1.651	D6S2725	0.432	5	0.007	−0.324	3
DXS1003	0.91	16	0.326	0.101	DXS10064	0.544	5	0.073	1.008	1.5
DXS1193	0.854	11	0.148	1.955	DXS10063	0.562	13	0.076	4.673	4.2
DXS6799	0.764	9	0.142	2.003	DXS10065	0.567	6	0.295	2.088	3.6
Average	0.815	11.0	0.225	1.204	Average	0.661	8.6	0.171	1.153	

NOTE.—H indicates heterozygosity; A indicates number of alleles.

genetic differences due to drift would be expected for X-linked loci, if the effective size is indeed smaller than that for autosomal loci. In addition, local selection involving recessive advantageous mutations would be expected to occur more frequently on the X chromosome (Charlesworth, Coyne, and Barton 1987), since recessive mutations at autosomal loci will be dominated by drift during their early history, until they have reached a high enough frequency for significant numbers of the homozygous recessive genotype to appear, at which point selection can act. By contrast, the phenotype of recessive advantageous mutations on the X chromosome will be immediately apparent in males, and hence selection will be much more effective. The fact that diversity is not lower for the X-linked loci, whereas genetic differentiation is larger for the X-linked loci (albeit not significantly so), suggests that differences in effective population size alone cannot explain these patterns. More loci need to be examined to see if genetic differentiation between human populations is indeed larger for X-linked loci, and (if so), to determine to what extent either a reduced effective population size or a greater propensity for local selection is responsible.

To detect candidate genomic regions under local selection, we calculated two measures of genetic distance for our data and looked for outliers (i.e., loci with unusually large genetic distance values). The rationale for this approach, as first proposed by Cavalli-Sforza (1966) and Lewontin and Krakauer (1973), is that since by definition local selection inflates allele frequency differences between populations, marker loci that show unusually large genetic distance values are good candidates for local selection. However, the extent to which an STR locus that is closely linked to a gene subjected to local selection will show an unusually large genetic distance value depends on a number of factors, including the strength of selection, the amount of time elapsed since selection began, the amount of recombination between the marker locus and the selected locus, the mutation rate for the STR locus, and the particular STR allele that was carried by the selected haplotype. These factors will have a different impact on R_{st} and ln RV values. Following a newly arisen favorable mutation, the frequency of the selected allele will increase, as will the frequency of the STR allele that is on the same haplotype as the selected allele. For ln RV values, the local selection will reduce the variance in the allele frequency distribution for the STR locus in the population in which the local selective sweep is occurring, relative to a population that is not experiencing the local selective sweep, leading to an unusually large ln RV value. Over time, new mutations at the STR locus, as well as recombination between the STR locus and the selected gene, will increase the variance in the allele frequency distribution. Eventually the variance in the selected and non-selected populations will equalize, and so the signal of local selection will not be evident in the ln RV value.

The impact of local selection on R_{st} values will depend on the particular STR allele that is on the haplotype of the selected allele. If the STR allele happens to be a common allele, then during the selective sweep R_{st} values will first increase moderately (due to the decrease in allelic variance) but then decrease as new mutations and/or recombination regenerate allelic variation at the STR locus. However, if the STR allele on the selected haplotype happens by chance to be a rare allele, then there will be a large increase in the R_{st} value, as the mode of the allele frequency distribution at the STR locus has shifted. The large R_{st} value will be maintained even as new

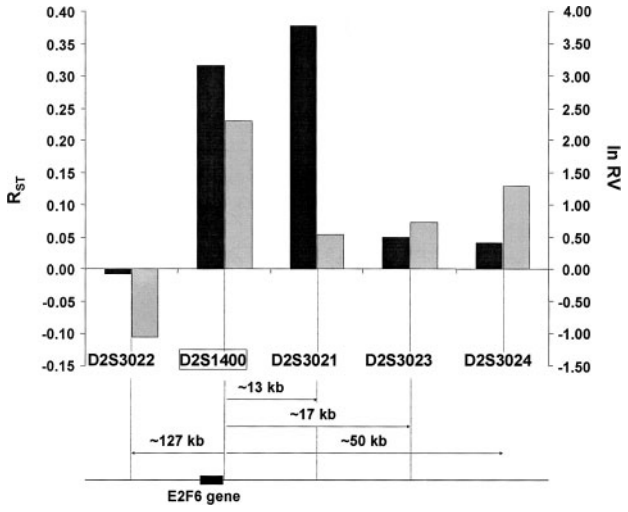

FIG. 6.—R_{st} (dark bars) and ln RV (light bars) values for additional STR loci in the vicinity of D2S1400. The gene for the E2F transcription factor 6 (E2F6) maps 0.9 kb downstream of the STR locus D2S1400 and spans a region of 20.6 kb.

Table 3
R_{st} and ln RV Values for Five Outlier Loci Also Analyzed by Rosenberg et al. (2002)

	This Study		Rosenberg et al. 2002[a]	
Locus	R_{st}	ln RV	R_{st}	ln RV
D16S539	0.268	−0.265	0.011	0.486
D21S1437	0.263	0.170	0.269	0.410
D9S2169	0.266	1.440	0.249	1.651
D2S1400	0.317	2.299	0.519	2.201
D6S1031	0.277	1.651	0.277	0.944
Average	0.278	1.059	0.265	1.138

[a] Comparing the French and Bantu samples.

alleles are generated by mutation, as the stepwise mutation process will generate new alleles around the new modal allele, and hence the entire allele frequency distribution at the STR locus will shift. Thus, the power of ln RV values to detect local selection will be highest immediately following the onset of selection and will then decline over time, whereas the power of R_{st} values should not decline with time, but instead will depend on the extent to which new modal alleles were produced as a consequence of selection.

We assumed that local selection would primarily influence Europeans, as modern humans originated in Africa, and hence new opportunities for local selection would have occurred as modern human populations spread out of Africa. Some support for this assumption comes from the distribution of ln RV values (fig. 3), in which there is an excess of extreme positive values (i.e., in the right-hand tail); since the variance in Africans appears in the numerator of the ln RV value, this indicates that there are many more loci showing significantly reduced variation in Europeans (relative to Africans) than in Africans (relative to Europeans). However, this should be interpreted cautiously, as an extreme bottleneck in Europeans, which is suggested by some genetic data (Tishkoff et al. 1996; Yu et al. 2002), could also lead to an excess of loci with significantly reduced variation in Europeans relative to Africans (Schlötterer 2002*a*).

Ideally, to identify significant outliers, the observed distribution of R_{st} and ln RV values would be compared with that expected under neutrality. Although some progress has been made toward understanding the statistical properties of R_{st} and ln RV values (Balloux and Goudet 2002; Schlötterer 2002*a*), the underlying assumptions of the models concerning demographic history, migration, and mutation raise questions as to the utility of using analytical approaches based on these models to identify loci in human populations that might be in genomic regions that have been subjected to local selection. A key issue, that has not been adequately addressed, is the extent to which lack of independence of population samples will influence the expected distribution of these statistics under neutrality versus selection (Nei and Chakravarti 1977; Nei, Chakravarti, and Tateno 1977). Further theoretical work on the statistical properties of R_{st} and ln RV values, as well as the development of new methods for detecting local selection (e.g., Sabeti et al. 2002) are required in order to make full use of the approach and the data.

We therefore adopted an empirical approach, based on the simple assumption that those loci with the largest R_{st} and/or ln RV values are the most likely candidates, and that if local selection has indeed been operating on the genomic regions containing such loci, then other loci in these genomic regions should also exhibit unusually large R_{st} and/or ln RV values. We found that this was indeed the case (table 2 and fig. 6); the additional STR loci that we characterized near the target loci had R_{st} and ln RV values that were significantly larger than average. However, even though the average values for the nearby loci were significantly larger, some of the individual nearby loci did not exhibit unusually large R_{st} and/or ln RV values. This could indicate that some of these genomic regions do not in fact exhibit unusually large genetic distances. Alternatively, the particular nearby STR locus may not have as much power as the original locus to detect unusually large genetic distances, as this will depend on the number of alleles and overall heterozygosity of the locus. Indeed, the average heterozygosity and number of alleles were both lower for the nearby loci than for the original target loci (table 2), significantly so for average heterozygosity (Mann-Whitney U test, $P < 0.01$) and nearly significantly so for the number of alleles ($P = 0.06$). Thus, the failure of a nearby locus to confirm the large R_{st} and/or ln RV value of a particular target locus does not rule out local selection on this genomic region.

Characterization of additional STR loci in the genomic region of interest, as was done for the region surrounding the E2F6 gene (fig. 6), may provide further information. E2F6 is a novel member of the E2F family of transcription factors, which regulate cellular proliferation and differentiation and are thought to play a role in cancer (Johnson and Schneider-Broussard 1998). E2F6 represses transcription and is a component of the mammalian polycomb complex (Trimarchi et al. 2001; Ogawa et al. 2002), suggesting that it plays a key role in normal developmental patterning. Although it is not obvious why local selection should have influenced this gene, characterizing the pattern of genetic variation at E2F6 should help address this question.

As a further check on the reproducibility of our results, we compared our results with another recent study that looked for evidence of selection on the human genome. Schlötterer (2002*a*) applied his ln RV statistic to 94 loci that had been typed in 10 African and non-African populations. Although the number of outliers was consistent with neutral expectations, he considered four loci as possible candidates for local selection. One of these four loci, D6S305, was also included in our study; Schlötterer (2002*a*) found highly reduced variation at this locus in African populations, which we also found (ln RV = –0.97, compared with the average ln RV value of 0.20). Additionally, Rosenberg et al. (2002) recently analyzed 377 STR loci in 1052 individuals from 52 populations. Five of the 11 loci that we identified with high R_{st} and/or ln RV values were also analyzed in their study. We obtained the data for the Bantu and French samples in their study (which would be most comparable to our African and European samples) and calculated R_{st} and ln RV values for these five loci (table 3). The mean R_{st}

and ln RV values were nearly identical for these five loci in both studies, and the mean R_{st} and ln RV values were significantly higher for these five loci in the Bantu/French comparison than for all of the loci in our study (Mann-Whitney tests: R_{st}, $Z = 2.89$, $P < 0.01$; ln RV, $Z = 2.82$, $P < 0.01$). Thus, loci that we identify as outliers in our study are also outliers when other samples are analyzed.

Although the genome scan approach identifies unusual genomic regions that differ reproducibly from the ''average'' genomic region, it does not distinguish between local selection versus some other explanation (such as extreme genetic drift) as the agent responsible for the unusual behavior of these genomic regions. Further characterization of interesting genes/DNA segments in these genomic regions, coupled with functional analyses of any relevant polymorphisms, is required to demonstrate conclusively that local selection has indeed influenced genetic variation in a particular genomic region. Nevertheless, genomic regions that show unusually large genetic differences between populations are obvious candidates for local selection (Schlötterer 2002*b*), and in this paper we have demonstrated that a genome scan is an effective means of identifying such genomic regions.

Further extensions to this approach include screening additional populations and incorporating additional loci into the genome scan (especially as developing technology enables genome scans based on SNPs). Moreover, loci under local selection in other species could also be identified by genome scans. Our results indicate that genome scans should aid in the identification of candidate regions under local selection in human populations, which will increase our knowledge of the selective factors and forces that have shaped human genetic variation.

Acknowledgments

We thank H. Soodyall and E. Edel for DNA and blood samples; I. Jonasson, A. S. Strand, and U. Gyllensten for genotyping; and W. Enard, M. Krawczak, S. Pääbo, M. Przeworski, A. Ryan, L. Vigilant, and G. Weiss for useful discussion. Supported by funds from the Max Planck Society.

Literature Cited

Allison, A. C. 1954. Protection afforded by sickle-cell trait against subterian malarial infection. Br. Med. J. **1**:290–294.

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

Baer, C. F. 1999. Among-locus variation in Fst: fish, allozymes and the Lewontin-Krakauer test revisited. Genetics **152**:653–659.

Balloux, F., and J. Goudet. 2002. Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Mol. Ecol. **11**:771–783.

Beaumont, M. A., and R. A. Nichols. 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. Roy. Soc. Lond. B Biol. Sci. **263**:1619–1626.

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

Caballero, A. 1995. On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics **139**:1007–1011.

Cavalli-Sforza, L. L. 1966. Population structure and human evolution. Proc. Roy. Soc. Lond. B Biol. Sci. **164**:362–379.

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94**:1041–1046.

Charlesworth, B., J. A. Coyne, and N. H. Barton. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130**:113–146.

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

Hill, A. V. S., C. E. M. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi, P. A. Rowe, S. Bennett, D. Brewster, A. J. McMichael, and B. M. Greenwood. 1991. Common West African HLA antigens are associated with protection from severe malaria. Nature **352**:595–600.

Johnson, D. G., and R. Schneider-Broussard. 1998. Role of E2F in cell cycle control and cancer. Frontiers Biosci. **3**:447–458.

Jorde, L. B., W. S. Watkins, M. J. Bamshad, M. E. Dixon, C. E. Ricker, M. T. Seielstad, and M. A. Batzer. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am. J. Hum. Genet. **66**:979–988.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. Genome Res. **12**:996–1006.

Lahr, M. M., and R. A. Foley. 1998. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. Yrbk. Phys. Anthropol. **41**:137–176.

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

———. 1975. Letters to the editors: testing the heterogeneity of *F* values. Genetics **80**:397–398.

Lindqvist, A. K., P. K. Magnusson, J. Balciuniene, C. Wadelius, E. Lindholm, M. E. Alarcon-Riquelme, and U. B. Gyllensten. 1996. Chromosome-specific panels of tri- and tetranucleotide microsatellite markers for multiplex fluorescent detection and automated genotyping: evaluation of their utility in pathology and forensics. Genome Res. **6**:1170–1176.

Luzzatto, L., E. Usanga, and S. Reddy. 1969. Glucose-6-phosphate dehydrogenase deficient red cells: resistance to infection by malarial parasites. Science **164**:839–841.

Miller, L. H., S. J. Mason, D. F. Clyde, and M. H. McGinniss. 1976. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. N. Engl. J. Med. **295**: 302–304.

Mountain, J. L. 1998. Molecular evolution and modern human origins. Evol. Anthropol. **7**:21–37.

Nei, M., and A. Chakravarti. 1977. Drift variances of Fst and Gst statistics obtained from a finite number of isolated populations. Theor. Popul. Biol. **11**:307–325.

Nei, M., A. Chakravarti, and Y. Tateno. 1977. Mean and variance of Fst in a finite number of incompletely isolated populations. Theor. Popul. Biol. **11**:291–306.

Nei, M., and T. Maruyama. 1975. Letters to the editors: Lewontin-Krakauer test for neutral genes. Genetics **80**:395.

Ogawa, H., K. Ishiguro, S. Gaubatz, D. M. Livingston, and Y. Nakatani. 2002. A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G_0 cells. Science **296**:1132–1136.

Payseur, B. A., A. D. Cutter, and M. W. Nachman. 2002. Searching for evidence of positive selection in the human

genome using patterns of microsatellite variability. Mol. Biol. Evol. **19**:1143–1153.

Robertson, A. 1975. Gene frequency distributions as a test of selective neutrality. Genetics **81**:775–785.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. 2002. Genetic structure of human populations. Science **298**:2381–2385.

Sabeti, P. C., D. E. Reich, J. M. Higgins et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature **419**:832–837.

Schlötterer, C. 2002*a*. A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160**:753–763.

———. 2002*b*. Towards a molecular characterization of adaptation in local populations. Curr. Opin. Genet. Dev. **12**: 683–687.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

Stoneking, M. 1993. DNA and recent human evolution. Evol. Anthropol. **2**:60–73.

Stoneking, M., J. J. Fontius, S. L. Clifford, H. Soodyall, S. S. Arcot, N. Saha, T. Jenkins, M. A. Tahir, P. L. Deininger, and M. A. Batzer. 1997. *Alu* insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res. **7**:1061–1071.

Tishkoff, S. A., E. Dietzsch, W. Speed et al. (15 co-authors). 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271**:1380–1387.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293**:455–462.

Trimarchi, J. M., B. Fairchild, J. Wen, and J. A. Lees. 2001. The E2F6 transcription factor is a component of the mammalian Bmi1-containing polycomb complex. Proc. Natl. Acad. Sci. USA **98**:1519–1524.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. Science **253**:1503–1507.

Vitalis, R., K. Dawson, and P. Boursot. 2001. Interpretation of variation across marker loci as evidence of selection. Genetics **158**:1811–1823.

Yu, N., F. C. Chen, S. Ota, L. B. Jorde, P. Pamilo, L. Patthy, M. Ramsay, T. Jenkins, S. K. Shyue, and W. H. Li. 2002. Larger genetic differences within Africans than between Africans and Eurasians. Genetics **161**:269–274.

Naruya Saitou, Associate Editor

Accepted January 20, 2003