

Population Differences in the Human Functional Olfactory Repertoire

Yoav Gilad and Doron Lancet

Department of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot, Israel

Olfactory receptors (OR) constitute the molecular basis for the sense of smell. They are encoded by a large multigene family that in humans includes approximately 400 functional genes and approximately 600 putative pseudogenes, distributed on all but two chromosomes. To examine the ethnogeographic variability in the functional chemosensory repertoire, we resequenced 32 OR loci reported to contain a single coding region disruption in seven Caucasians and seven Pygmies. Thirteen of the 32 OR loci were found to have an interrupted coding region in all 28 alleles sampled, seven had an intact form in all the individuals examined, and 12 were polymorphic, segregating both the intact and the null variants. Among the latter loci, the frequency of the null allele was higher in Caucasians than in Pygmies, suggesting that African populations may have a larger repertoire of functional OR genes. Interestingly, when analyzing the entire OR coding regions, we find an excess of high-frequency derived alleles at many loci in the Caucasian sample but less so in the Pygmy sample. Our observations are unlikely to be accounted for by simple demographic models but may be explained by positive selection acting on OR loci in Caucasians.

Introduction

With the elucidation of the sequence of the human genome (Lander et al. 2001) a key remaining problem is to identify polymorphisms in human populations and in particular to identify functionally important variation. One approach to discovering functional variants is to use linkage-disequilibrium mapping to identify genes underlying particular phenotypes (Risch et al. 1996). Another approach is to use statistical analyses of DNA sequence variation to try to identify regions that have recently been targets of positive natural selection (Tishkoff et al. 2001; Gilad et al. 2002; Hamblin et al. 2002). This second approach has also been useful in identifying population-specific effects of selection, including differences between African and non-African populations (e.g., the Duffy locus [Hamblin et al. 2002]; the Mc1r locus [Harding et al. 1997]).

Olfactory receptor (OR) genes form the genetic basis for the sense of smell (Buck and Axel 1991; Lancet and Ben-Arie 1993; Mombarts 1999). There are more than 1,000 ~1 kb intronless OR coding regions in the human genome (Glusman et al. 2000; Glusman et al. 2001; Zozulya et al. 2001). These are found in gene clusters across all human chromosomes except 20 and Y.

One of the most surprising attributes of the human olfactory repertoire is its vast degree of pseudogenization (Glusman et al. 2001; Zozulya et al. 2001; Zhang and Firestein 2002). More than 60% of the OR genes bear one or more coding region disruptions that likely result in a functional inactivation of the coded protein. This appears to be a rather recent evolutionary phenomenon, since mouse has a similar repertoire size but a much smaller pseudogene fraction (~20%) (Young et al. 2002; Zhang and Firestein 2002). A graded process of gene inactivation seems to have taken place in mammalian evolution (Rouquier, Blancher, and Giorgi 2000), with an especially steep decline in primates (Sharon et al. 1999). These findings may be explained by the transition from macro-smatic mammals (e.g., mouse, rat, and dog), with a highly

acute and behaviorally relevant sense of smell, to micro-smatic apes, in whom vision and audition are much more significant than olfaction.

Of the ~600 OR pseudogenes in the human genome, there are 67 whose coding region is interrupted in only one position (Glusman et al. 2001); these are likely to be the result of recent mutations. They may also be promising candidates for polymorphisms that affect olfactory function. It has been argued (Lancet et al. 1993b; Firestein 2001) that such genetic polymorphisms could at least partly explain the observed interindividual variation in olfactory sensitivity (Whissell-Buechy and Amoore 1973; Amoore 1977; Wysocki and Beauchamp 1984; Gross-Isseroff et al. 1992).

We have previously reported that in a cluster on chromosome 17, intact OR genes appear to be evolving under positive selection (Gilad et al. 2000). We also characterized pseudogenes in this cluster and showed two of them to be segregating gene and pseudogene variants (Menashe et al. 2002). The present study describes the resequencing of 32 single-disruption pseudogenes belonging to 14 OR gene clusters in 14 human individuals from two population groups. We find that the two human populations appear to have significant differences in the size of the intact olfactory repertoire. They also appear to differ in the patterns of genetic variation of the entire coding region, potentially reflecting distinct selective pressures.

Materials and Methods

DNA Samples

Human genomic DNA was from two sources: (1) Genomic DNA of two unrelated Ashkenazi Jews isolated from whole blood using the Genomix DNA preparation kit (Talent SRL, Trieste Italy) and (2) seven DNA samples of Pygmy individuals and five samples of Europeans purchased from Corriel Cell Repositories. The European and the Ashkenazi Jew samples are pooled and referred to as Caucasian. Genomic DNA from two chimpanzees (*Pan troglodytes*) was isolated from whole blood (kindly provided by Dr. Yigal Horvitz from the Israeli Safari Zoo, Ramat Gan) using the Genomix DNA preparation kit.

Key words: olfactory receptors, positive selection.

E-mail: yoav.gilad@weizmann.ac.il.

Mol. Biol. Evol. 20(3):307–314, 2003

DOI: 10.1093/molbev/msg013

© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

OR Loci

The OR pseudogenes were identified in the HORDE database (<http://bioinformatics.weizmann.ac.il/HORDE/>), which is based on sequences mined from the public database (Glusman et al. 2001). HORDE contains a conceptually translated protein for every pseudogene, according to an OR genes alignment, which is specific to every OR gene family (Glusman et al. 2001). We chose OR loci that contain only a single disruption in their coding region within the first 290 amino acids. It should be pointed out that our definition of a putatively functional OR gene is based only on the coding region. Other reasons (such as mutations in the promoter region) may indeed cause a gene with an intact coding region not to be functional. Future OR gene expression studies will address this issue. In this study, we refer to OR loci with an intact coding region as “intact” and to OR loci with disrupted coding region as “pseudogenes.”

PCR Procedures

Primers for PCR amplification and for sequencing were designed to amplify the full open-reading frame of the 32 human OR genes. PCR was performed in a total volume of 25 μ l, containing 0.2 μ M of each deoxynucleotide (Promega, Madison, Wis.), 50 pM of each primer, and PCR buffer containing 1.5 mM MgCl₂, 50 mM KCl, 10 mM Tris pH 8.3, 1 unit of Taq DNA polymerase (Boehringer, Mannheim, Germany), and 50 ng of genomic DNA. PCR conditions were as follows: 35 cycles of denaturation at 94°C, annealing at 55°C, and extension at 72°C, each step for 1 min. The first step of denaturation and the last step of extension were 3 min and 10 min, respectively. PCR products were separated on a 1% agarose gel to view their size, and they were purified using the High Pure PCR Product Purification Kit (Boehringer, Mannheim, Germany).

DNA Sequencing

Sequencing reactions were performed on PCR products or clones in both directions with a dye terminator cycle sequencing kit (PerkinElmer, Wellesley, Mass.) on an ABI 3700 automated sequencer (PerkinElmer, Wellesley, Mass.). The OR coding regions were sequenced from both ends for each individual. After base calling with the ABI Analysis Software (version 3.0), the data were edited and assembled using the Sequencher program (version 4.0) (GeneCodes Corp. Ann Arbor, Mich.), and DNA polymorphisms were thus identified. After the assembly of each species separately, the chimpanzee sequences were added to the human assembly to infer divergent sites.

Statistical Analyses

We calculate three summaries of diversity levels: Watterson's θ_W (Watterson 1975), based on the number of segregating sites in the sample; π (Nei and Li 1979), the average number of pairwise differences in the sample; and θ_H (Fay and Wu 2000), a summary that gives more weight to high-frequency derived variants. Under the standard

neutral model of a random-mating population of constant size, all three summaries estimate the population mutation parameter $\theta = 4N\mu$, where N is the diploid long-term inbreeding effective population size and μ is the mutation rate per generation. To test whether the frequency spectrum of mutations conformed to the expectations of this standard neutral model, we calculate the value of two test statistics: Tajima's D statistic (Tajima 1989b), which considers the difference between π and θ_W , and Fay and Wu's H -test (Fay and Wu 2000), which considers the difference between π and θ_H .

We ran coalescent simulations of an infinite-site locus with no recombination, assuming a constant population size and random-mating. To estimate the P -values for H and D , we tabulated the proportion of simulated data sets with an $H(D)$ value more extreme (i.e., more negative) than the observed one. Both D and H were used as one tailed “tests of neutrality.” When analyzing the subset of ORs that segregate between gene and pseudogene variants, we can no longer use standard coalescent assumptions. Indeed, an ascertainment bias is introduced by picking those ORs that were pseudogenes in a sample size of one (the reference database). Since pseudogenes always appear to be the derived state (see below), this is equivalent to sampling loci with a derived allele in a sample size of one. Note further that the genealogies of sites within a locus are correlated. Thus, if there is one allele at higher frequency than expected at random, it is more likely that alleles at nearby sites will also be at slightly higher frequencies. As a result, this set of ORs is more likely to contain one or more alleles that are at high frequency than if they were chosen at random. This might bias the results of tests such as D and H . To correct for this, we generated simulated data sets for 28 chromosomes and considered the first individual as simulating the sequence from the original database. If this individual had exactly one derived allele in a defined segment L of its chromosome, we retained the sample of 28 chromosomes. Otherwise, we rejected the entire sample. L was chosen so that the probability that an individual will have one derived allele in that segment is equal to the estimated proportion of single disruption OR pseudogenes that are segregating gene/pseudogenes in our sample (~ 0.3).

Estimating Recombination Rates

The recombination rate per generation was estimated for each cluster using the data of Payseur and Nachman (2000) (see <http://eebweb.arizona.edu/nachman/publications/data/microsats.html>), which is based on a comparison of a physical map (the GB4 radiation hybrid map) and the Genethon genetic map (see Payseur and Nachman 2000 for details). To estimate the rate for this locus, we used the closest microsatellites for each cluster, according to a Blast search of the corresponding OR gene sequence using NCBI's map viewer. This method effectively yields estimates of the rate of crossing-over alone, since gene conversion contributes little to the rate of gamete exchange for markers that are far apart (~ 1 Mb) (Pritchard and Przeworski 2001).

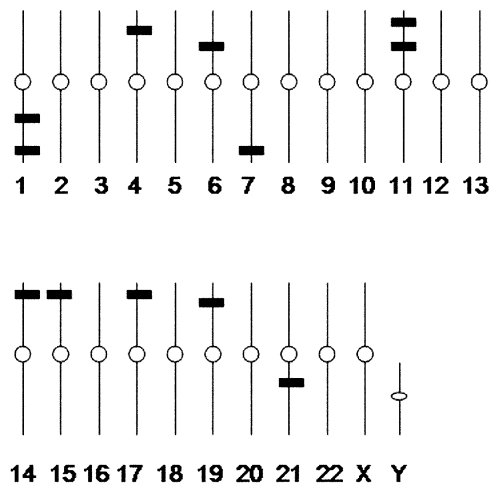


FIG. 1.—Location of the chosen OR genes. A heuristic drawing of human chromosomes. Circles indicate centromeres and black rectangles indicate the locations of OR clusters from which the 32 OR genes were sampled (see table 1 for specific location).

Results

Segregating Pseudogenes

We selected 32 single-disruption OR loci for the analysis. These are located in 14 different clusters on 10 different chromosomes (fig. 1). The 1 kb coding regions were resequenced in 14 human individuals (28 alleles) comprising seven Pygmy Africans and seven Caucasians. ORs from two chimpanzees were also similarly analyzed in order to determine the ancestral allele. We could not amplify the chimpanzee orthologous gene for the human OR2A9P, and therefore this locus was excluded from any subsequent analysis that required outgroup sequence information. All sequences were deposited to GenBank (accession numbers AF546191–AF546699).

Of the 32 OR loci, seven were found to have an intact open-reading frame (ORF) in all 14 human individuals and in two chimpanzees. Such ORs may have been incorrectly annotated in the database due to sequencing errors or may constitute rare pseudogene polymorphisms. Thirteen other loci had interrupted ORFs in all humans in our sample, all with the same disruptions as in the corresponding HORDE reference sequences. Six of these 13 loci were also found to have the disrupted coding region in the four chimpanzee alleles, whereas the disruptions in the other seven loci were specific to humans.

We observed 12 loci with a polymorphic disruption, that is, coding regions for which the intact and the disrupted versions (null allele) were segregating in the human sample. These loci are henceforth referred to as segregating pseudogenes, or SPGs. Five of the segregating disruptions are nonsense mutations and seven are frameshift mutations that lead to a premature stop codon. In 11 cases, the null variant is inferred to be the derived state from the observation that the orthologous gene was intact in all the chimpanzee sequences. For one human pseudogene (OR4Q1P), we found the same coding region disruption also in the four chimpanzee alleles. This might be a shared polymorphism in humans and chimpanzees.

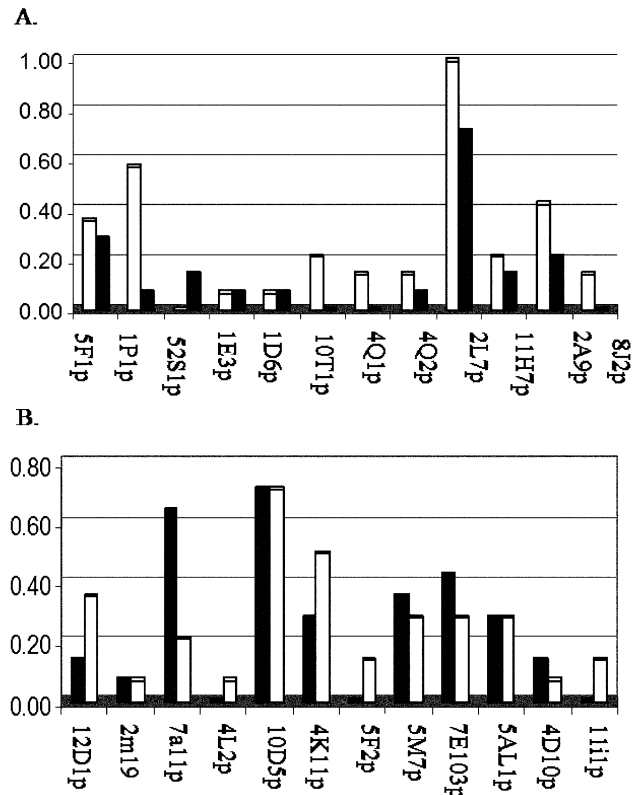


FIG. 2.—Population differences in the frequency of alleles. The values for Caucasians are plotted as the dark bars and those for the Pygmies as empty bars. OR genes are indicated by their official Human Genome Organization human gene nomenclature symbols (Glusman et al. 2000). A, Population differences in the fraction of the intact OR allele. B, Population differences in the fraction of derived alleles that were identified in a randomly chosen Caucasian individual.

Yet, since humans and chimpanzees split approximately $25N$ generations ago (where N is the effective population size), this is more likely to be the result of recurrent mutation than the persistence of ancestral polymorphism (Halushka et al. 1999).

We examined the frequency of the intact OR alleles for the 12 SPGs. The null allele frequency for two SPGs was equal in the two populations (OR1E3p and OR1D6p). For nine out of the other 10 SPGs, the null allele was found to be at a higher frequency in Caucasians than in Pygmies (fig. 2A; two-tailed sign test, $P < 0.021$). The frequencies of the null alleles are on average 13% higher in the Caucasians compared with the Pygmies. Similarly, while 13 disruptions are fixed in both populations (table 1), there is one additional fixed disruption in the Pygmies (OR52S1), compared with three additional fixed disruptions in Caucasians (OR10T1p, OR4Q1p, and OR8J2p). Thus, in our sample, the intact OR gene repertoire of Caucasians appears to be smaller than that of the Pygmies.

The OR sequences in the human genome databases derive from several ethnogeographical groups, but the ethnic origin of particular OR pseudogene sequences used here is unknown. If most such sequences are of Caucasian origin, then the finding of a higher frequency of pseudogenes in Caucasians relative to Pygmies may simply

Table 1
Data for the 32 Chosen OR Loci

ALL	Pseudo?	Position	Length	S ^a	π %	θ per bp %	Tajima's <i>D</i> ^b	<i>H</i>	<i>P</i> (<i>H</i>)	<i>r</i> ^c
52m1p	n	11–3.04Mb	952	3	0.10	0.11	–0.21	0.10	0.33	1.29
10j4p	n	unknown	660	3	0.09	0.12	–0.80	0.43	0.52	nd
4s2p	n	11–50.99Mb	934	1	0.01	0.03	–0.74	0.12	0.28	1.50
3a8p	n	5–?	648	2	0.08	0.08	–0.07	–0.56	0.12	nd
5aL1p	n	11–52.13Mb	919	2	0.05	0.08	–1.00	–1.55	0.04	1.50
4d10p	n	11–55.86Mb	911	3	0.06	0.08	–0.75	–1.31	0.06	0.71
2a12p	n	7–148.33Mb	466	0	0.00	0.00	nd	nd	nd	1.00
51f1p	seg	11–3.60Mb	970	7	0.33	0.21	1.70	–1.68	0.20	1.29
1p1p	seg	17–2.99Mb	993	6	0.23	0.16	1.35	–0.73	0.33	1.70
52s1p	seg	11–3.79Mb	964	4	0.13	0.13	–0.12	–1.73	0.15	1.29
1e3p	seg	17–2.99Mb	948	5	0.16	0.14	0.52	–1.23	0.21	1.70
1d6p	seg	11–35.02Mb	642	5	0.22	0.20	0.33	–3.78	0.06	0.96
10t1p	seg	1–154.34Mb	922	4	0.14	0.14	–0.02	–4.34	0.05	0.87
4q1p	seg	15–1.66Mb	934	4	0.11	0.13	0.01	–1.34	0.21	0.79
2L7p	seg	1–254.55Mb	380	3	0.22	0.20	0.22	–1.20	0.20	0.94
4q2p	seg	14–0.08Mb	891	5	0.20	0.14	1.03	0.44	0.50	3.63
11h7p	seg	14–0.33Mb	962	5	0.16	0.13	0.63	–2.65	0.24	3.64
2a9p	seg	7–149Mb	993	2	0.06	0.08	–0.61	nd	nd	1.28
8j2p	seg	11–51.94	954	3	0.07	0.08	–0.40	–1.23	0.18	1.50
7e103p	y	4–5.55Mb	651	5	0.35	0.24	1.44	0.22	0.37	1.28
5f2p	y	11–51.76Mb	943	4	0.08	0.11	–0.60	–4.34	0.01	1.50
4k11p	y	21–8.12Mb	967	4	0.23	0.13	2.03	–1.61	0.07	1.76
7a11p	y	19–17.72Mb	936	3	0.11	0.08	0.72	–0.73	0.15	0.55
4a8p	y	11–50.28Mb	924	3	0.06	0.08	–0.63	0.39	0.49	1.50
12d1p	y	6–33.02Mb	929	3	0.04	0.08	–1.13	–3.03	0.02	1.10
5m7p	y	11–52.17Mb	935	1	0.06	0.05	0.21	–0.20	0.23	1.50
4L2p	y	14–0.21Mb	918	2	0.06	0.06	0.13	–0.31	0.20	2.59
5an2p	y	11–55.69Mb	936	1	0.01	0.03	–0.74	0.12	0.33	0.71
4c1p	y	11–50.93Mb	926	5	0.02	0.02	–0.18	0.65	0.67	1.50
10d5p	y	11–138.10Mb	913	4	0.11	0.11	–0.14	–1.24	0.09	1.78
11i1p	y	1–126.31Mb	933	5	0.18	0.14	0.84	–2.65	0.03	0.90
4a13p	y	11–50.81Mb	945	0	0.00	0.00	nd	nd	nd	1.50

^a The number of SNPs found for each locus.^b Significant Tajima's *D* values ($P < 0.05$) are bolded.^c Recombination rate (cM/Mb) for this genomic region as reported from a comparison of genetic and physical maps (see *Materials and Methods*).

reflect an ascertainment bias. In order to test this possibility, we repeated the same analysis with the loci that were fixed as genes or pseudogenes in our sample, assuming that selecting the derived allele in a single Caucasian individual mimics the original process of studying single-disruption pseudogenes from the database. We thus picked an individual at random from the Caucasian sample and compared the frequency of the derived allele of the first SNP in the coding region between the two population samples. We found that the derived alleles were not significantly often at higher frequency in the Caucasian sample (fig 2B; two-tailed sign test, $P = 0.77$). The average frequency of the derived allele in Caucasians is only 1% higher than in the Pygmies; this difference is significantly

smaller than the 13% difference observed for the sites that create SPG (t -test, $P < 0.043$). Thus, ascertainment bias does not appear to form the basis for our SPG observations.

Positive Selection

We also examined the polymorphism data from the entire OR coding regions. A total of 107 single nucleotide polymorphisms (SNPs) were observed, appearing in all but two of the ORs (tables 1 and 2); 19 were seen only once in the sample (singletons). The majority of the SNPs were found in both population groups, whereas 11 were specific to the Pygmies and four to the Caucasians (not including singletons).

Table 2
Statistical Properties of the Total Sample and the Two Population Groups

	<i>N</i>	bp ^a	S ^b	θ per bp %	π %	<i>H</i>	<i>P</i> (<i>H</i>)	Tajima's <i>D</i>	<i>P</i> (<i>D</i>) ^c
Total sample	28	27,899	107	0.09	0.09	–32.38	0.04	0.10	0.62
Pygmies	14	27,899	94	0.09	0.10	–21.23	0.11	0.32	0.82
Caucasians	14	27,899	76	0.07	0.07	–30.24	0.02	–0.13	0.56

^a The total number of base pairs sequenced.^b The number of SNPs that were found.^c *P*-values are reported for multiple-loci Tajima's *D* statistic.

We have previously reported that positive selection is acting on intact OR genes on human chromosome 17 (Gilad et al. 2000). Consistent with previous results, we observe reduced number of polymorphic sites in intact genes compared with pseudogenes: 2.00 ± 1.5 and 3.73 ± 1.6 , respectively (t -test, $P = 0.015$). We calculated two statistics of the frequency spectrum to test for deviation from neutral expectations: Tajima's D statistic (Tajima 1989b), and Fay and Wu's H -test (Fay and Wu 2000). We assessed the P -values of the H -test and Tajima's D statistic using coalescence simulations with no recombination, correcting for the ascertainment bias in the SPGs (see *Materials and Methods*). For the intact genes and fixed pseudogenes, the expectation for both tests under the standard neutral model is zero. Although the average Tajima's D value is slightly positive for our entire data set (table 2), it is negative on average for the intact OR genes (table 1), consistent with our previous observation (Gilad et al. 2000).

Fay and Wu's H -test is significant for five out of 29 (17.2%) of the OR loci at the 5% level. The test is also significant if recombination between loci is ignored and all the ORs are (conservatively) treated as one completely linked locus (table 2, $P < 0.04$). The H -test finding reflects a relative excess of high-frequency derived alleles in our data. This again, is consistent with a model where positive natural selection has elevated the frequencies of alleles at linked sites (Fay and Wu 2000).

Population-Specific Positive Selection

We performed SNP analyses on each population group separately. Variability values were significantly higher (see table 2) for the Pygmies compared with the Caucasians (t -test, $P < 0.01$), consistent with results of other studies of African and non-African populations (Nickerson et al. 1998; Frisse et al. 2001). Although Pygmies have higher variability values, the number of singletons is practically identical in the two population groups. Accordingly, the average Tajima's D value for the entire data set is lower in Caucasians (table 2), although none of the Tajima's D values are significantly different than zero (see table 2 for P -values for multiple-loci test). We also performed Fay and Wu's H -test (Fay and Wu 2000) for separate loci (see P -values in fig. 2) and for the combined data set for each population (table 2). Whereas only three loci (10%) had a significant H -test at the 5% level in the Pygmy sample (OR4D10p, OR5F2p, and OR12D1p), nine loci (33%) were significant for the Caucasians (OR4D10p, OR1E3p, OR1D6p, OR10T1p, OR4Q1p, OR5F2p, OR7A11p, OR10D5p, and OR11i1p). Notably, in 20 out of 26 (77%) possible pairwise comparisons between the two populations, the H -test P -value was smaller for the Caucasians (fig. 3; sign test, $P < 0.009$). The H -test for the entire data set (table 2) is significant for the Caucasians ($P < 0.02$) but not for the Pygmies ($P < 0.11$). This indicates that there are more high-frequency derived alleles in Caucasians than in Pygmies and points to positive selection acting on the olfactory repertoire in Caucasians.

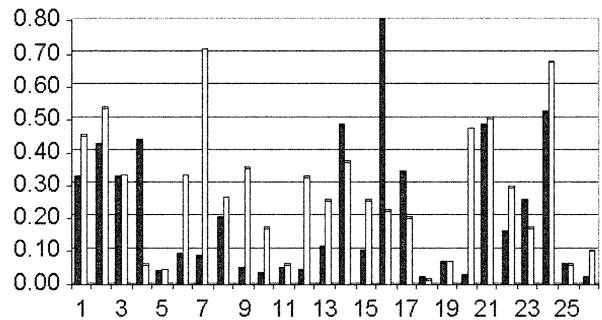


FIG. 3.— H -test P -values for Pygmies and Caucasians. Some comparisons were not possible since no SNPs were found for one of the population groups. The values for Caucasians are plotted as the solid bars and those for the Pygmies as empty bars. The H -test P -values are given for all the 26 OR loci for which SNPs were found in both population groups (ordered according to table 1).

A previous study of a single OR gene cluster that sampled only Caucasians found reduced levels of polymorphism in genes relative to divergence levels, as expected from models of positive selection (Gilad et al. 2000). Since the ORs we consider in this study are found in many different clusters, explanations that invoke positive selection would require multiple selective sweeps in different regions of the genome. Under a model of recurrent selective sweeps affecting diversity levels at ORs, we would expect ORs in regions of higher recombination to harbor higher levels of diversity (Maynard-Smith and Haigh 1974; Braverman et al. 1995). To test this hypothesis, we compared estimates of the recombination rate (table 1) and levels of nucleotide diversity among loci by binning loci into regions of high and low recombination rates (greater or less than the sample mean of 1.3 cM/Mb). In Caucasians but not in Pygmies, the mean nucleotide diversity value is significantly higher for loci in regions of high recombination relative to regions of low recombination (one-tailed, $P = 0.04$ and $P = 0.10$, respectively).

Discussion

Allele Frequency Differences Between Populations

Only a few cases of segregating OR gene disruptions have been previously reported (Younger et al. 2001; Menashe et al. 2002). Here we describe 12 segregating pseudogenes (SPGs) throughout the human genome, of which 10 are novel. We found higher frequencies of the intact alleles in the Pygmies compared with the Caucasians. Assuming that ORs with intact coding regions are indeed functional, in our sample the Caucasians have a smaller functional OR repertoire size. An analysis of randomly chosen sites in non-SPG loci indicated that ascertainment bias is not likely to be the reason for this observation. Moreover, a study of three SPGs on chromosome 17, not selected based on any prior knowledge, found a higher frequency of intact alleles in Africans (Menashe et al. 2002). The same was also true for two SPGs that were identified in the database as intact genes (Y. Gilad, I. Menashe, and D. Lancet, unpublished results). In summary, in five out of five cases chosen

without an ascertainment bias, the same pattern is observed (two-tailed sign test, $P = 0.063$). Thus, there appear to be differences in the functional olfactory repertoire size between Pygmies and Caucasians. Whether this points to a general difference between populations with different evolutionary histories remains to be investigated.

Can Simple Demographic Models Explain Our Observations?

Humans have experienced population growth, a process expected to lead to a skew in the frequency spectrum toward rare alleles (Tajima 1989a; Slatkin and Hudson 1991). Consistent with this theoretical expectation, several studies of nucleotide variation in humans have reported negative Tajima's D values (Nachman 2001; Stephens et al. 2001). Our observation of overall slightly positive Tajima's D values (table 2) is inconsistent with such a scenario. Moreover, under a neutral model with population growth, high-frequency derived alleles are less likely than under a constant population size model (Przeworski 2002), yet the H statistic is sharply negative at many loci in our Caucasian sample, indicating an excess of high-frequency derived alleles and suggesting that a simple model of population growth cannot account for the allele frequency distribution reported here.

The observation of lower diversity and relative paucity of rare alleles in non-African samples has led researchers to suggest that such populations have experienced bottlenecks (Przeworski, Hudson, and Di Rienzo 2000). Consistent with this model, we observe a lower diversity level in the Caucasian sample. Furthermore, it is known that the H -test can reject at higher than the nominal rejection probability under some recent bottleneck models (Przeworski 2002). However, we do not observe fewer rare alleles in the Caucasian sample compared with the Pygmies, as evidenced by the lower Tajima's D value (table 2) and the observed higher proportion of singletons, so a simple bottleneck also seems an unlikely explanation for our observations.

Although we studied only one African population (Pygmies), we used individuals from several Caucasian populations. A demographic model of structured populations might account for the significant H -test if the different demes were sampled very unequally (Przeworski 2002). However, a population structure model is not expected to lead to decreased variability in Caucasians. Most importantly, neither this nor any other simple demographic model can account for the observation of higher frequency of pseudogene alleles in Caucasians compared with the Pygmies.

Purifying Selection as an Explanatory Model

An explanation for our observation of a greater functional olfactory repertoire size in Pygmies may be population-specific purifying selection. If in the Pygmies there is a greater constraint to keep the ORs intact than in the Caucasians, this might account for the observed differences. Yet, purifying selection in Pygmies but not

in Caucasians would result in lower variability values in Pygmies compared with Caucasians and in a skewed frequency spectrum towards rare alleles in the Pygmies but not in Caucasians (Akashi 1999), inconsistent with our findings. Further, purifying selection could not account for the high proportion of high-frequency derived alleles in our sample.

Nevertheless, if we assume high constraint in Pygmies, specifically to maintain an intact coding region, but allow vast changes in amino acid composition, the previous reservations would be withdrawn. Yet, under such a model, one would not expect the Pygmies to have the pseudogene alleles at high frequency, in contrast to our previous (Menashe et al. 2002) and current observations (fig. 2A).

Positive Selection

An alternative to demographic explanations or purifying selection might be population-specific positive selection. Selective sweeps at OR loci in Caucasians but not in Pygmies would be consistent with both the observation of a higher proportion of high-frequency derived alleles in the Caucasians than in the Pygmies and a stronger positive correlation between recombination rate and variability values in Caucasians than in Pygmies.

It should be noted that one of our observations is not expected under a simple model of selective sweeps in Caucasians. Under such a model, we would predict a reduction of variability in Caucasians compared with Pygmies beyond what is normally found at neutrally-evolving regions. In a study of noncoding regions in 15 Hausa Africans and 15 Italians, variability levels in the latter sample were approximately 30% lower than in the former (Frisse et al. 2001). The decrease in variability in Caucasians relative to Pygmies observed in the present study is not statistically different from such neutral predictions (results not shown).

One explanation might be that the selective sweeps in the OR clusters are still ongoing. If the sweeps occurred in the time since the common ancestor of both populations, many of these sweeps will not be complete. Under simplifying assumptions, the time from introduction to fixation of a strongly favored allele is approximately $2\log(2N)/s$ (where N is the effective population size and s the selection coefficient) (Stephan, Thomas, and Lenz 1992). For $N = 10^4$ and $s = 1\%$, this duration is approximately 2000 generations, or 40,000 years, approximately the time since the bottleneck in Caucasian populations estimated by Reich et al. (2001). Incomplete sweeps may explain our observation of many significant H -tests without a significant excess of rare alleles (non-significant Tajima's D values).

There are several known differences in selective pressures between sub-Saharan African and non-African populations (Harding et al. 2000; Hamblin, Thompson, and Di Rienzo 2002). In a well-studied example (Hamblin, Thompson, and Di Rienzo 2002), the Duffy blood group locus, both reduced variability levels and a significant H -test were observed in the African (Hausa) sample. This

confirmed a prior hypothesis for ethnogeographically-restricted positive selection related to exposure to the malaria agent *Plasmodium vivax*. The signature of positive selection was not as clear-cut in the non-African populations (Chinese and Italians), and the patterns of diversity could not be accounted for by a simple selective sweep model. The possible interaction of selection with population history created a complex pattern, which was hard to interpret even for a single locus. For OR loci, epistatic effects would further complicate the model.

Olfactory Receptor Evolution

Olfactory receptors are the largest gene family in the genome (Glusman et al. 2001). These receptors do not act as individual entities. Rather, the perception of a single odor probably requires more than one OR (Lancet et al. 1993a; Hildebrand and Shepherd 1997; Malnic et al. 1999), and every OR is capable of binding more than one odorant. This suggests that mutations that disrupt an OR gene may be allowed to drift to high frequencies as long as the given genome carries another intact OR whose odorant-binding function is sufficiently similar. More appreciable selective constraint may emerge only once this condition is no longer fulfilled.

We found differences in the proportion of intact OR coding regions between Pygmies and Caucasians, as well as differences in the frequency spectrum. If Caucasian populations lost a subset of olfactory binding abilities, for example as the result of a recent bottleneck as proposed by (Reich et al. 2001), compensatory mutations that restore the olfactory repertoire could then have been favored. These would be mutations in intact OR genes that change the function of the receptor. Alternatively, positive selection on ORs in this non-African population may result from adaptation to new environments. Chemosensory ligand specificity appears to rest in a relatively small number of complementarity-determining residues (CDRs) (Pilpel and Lancet 1999). Thus, a small number of mutations could allow a functionally-dispensable OR gene to assume one of the OR specificities lost by pseudogene fixation.

If the advantageous mutations occur linked to a null allele, hitchhiking due to favorable mutations will tend to augment the null allele frequency. If one assumes that the selective sweeps acted on mutations that compensated for lost olfactory capabilities in Caucasians, it is probable that these mutations would occur in the same genetic cluster as the pseudogene. Indeed, ORs that belong to the same family or subfamily are more likely to reside in the same genomic cluster (Glusman et al. 2001; Young et al. 2002).

A simple model of positive selection can account for most but not all of our findings. Such incomplete agreement may not be unexpected, as existing models of selection do not take into account salient features of the biology of OR genes, such as the epistatic relations between the genes. Teasing apart the demographic and selective factors will require neutral reference loci surveyed in the same individuals.

Acknowledgments

D.L. holds the Ralph and Lois Silver Chair in Human Genomics. This work was supported by the Crown Human Genome Center at the Weizmann Institute of Science, by the Alfried Krupp foundation, Germany, and by an Israel Ministry of Science grant to the National Laboratory for Genome Infrastructure. We thank Molly Przeworski for providing a simulation to correct for ascertainment bias and Jody Hey for the multiple-loci Tajima's *D* statistic program. We are grateful to Michael Nachman, Molly Przeworski, and Jeff Wall for comments on the manuscript.

Literature Cited

- Akashi, H. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**:221–238.
- Amoore, J. E. 1977. Specific anosmia and the concept of primary odors. *Chem. Senses Flavor* **2**:267–281.
- Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**:783–796.
- Buck, L., and R. Axel. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**:175–187.
- Fay, J. C., and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405–1413.
- Firestein, S. 2001. How the olfactory system makes sense of scents. *Nature* **413**:211–218.
- Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack, and A. Di Rienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**:831–843.
- Gilad, Y., S. Rosenberg, M. Przeworski, D. Lancet, and K. Skorecki. 2002. Evidence for positive selection and population structure at the human MAO-A gene. *Proc. Natl. Acad. Sci. USA* **99**:862–867.
- Gilad, Y., D. Segre, K. Skorecki, M. W. Nachman, D. Lancet, and D. Sharon. 2000. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nat. Genet.* **26**:221–224.
- Glusman, G., A. Bahar, D. Sharon, Y. Pilpel, J. White, and D. Lancet. 2000. The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mamm. Genome* **11**:1016–1023.
- Glusman, G., I. Yanai, I. Rubin, and D. Lancet. 2001. The complete human olfactory subgenome. *Genome Res.* **11**:685–702.
- Gross-Isseroff, R., D. Ophir, A. Bartana, H. Voet, and D. Lancet. 1992. Evidence for genetic determination in human twins of olfactory thresholds for a standard odorant. *Neurosci. Lett.* **141**:115–118.
- Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**:239–247.
- 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.
- Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J.

- Cox, J. A. Schneider, D. S. Moulin, and J. B. Clegg. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**:772–789.
- Harding, R. M., E. Healy, A. J. Ray et al. 2000. Evidence for variable selective pressures at MC1R. *Am. J. Hum. Genet.* **66**:1351–1361.
- Hildebrand, J. G., and G. M. Shepherd. 1997. Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu. Rev. Neurosci.* **20**:595–631.
- Lancet, D., and N. Ben-Arie. 1993. Olfactory receptors. *Current Biology* **3**:668–674.
- Lancet, D., N. Ben-Arie, S. Cohen et al. 1993a. Olfactory receptors: transduction, diversity, human psychophysics and genome analysis. *Ciba Found. Symp.* **179**:131–141.
- Lancet, D., R. Gross-Isseroff, T. Margalit, E. Seidmann, and N. Ben-Arie. 1993b. Olfaction: from signal transduction and termination to human genome mapping. *Chem. Senses* **18**:217–225.
- Lander, E. S., L. M. Linton, B. Birren et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Malnic, B., J. Hirono, T. Sato, and L. B. Buck. 1999. Combinatorial receptor codes for odors. *Cell* **96**:713–723.
- Maynard-Smith, J. M., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**:23–35.
- Menashe, I., M. Orna, D. Lancet, and Y. Gilad. 2002. Population differences in haplotype structure within a human olfactory receptor gene cluster. *Hum. Mol. Genet.* **11**:1381–1390.
- Mombarts, P. 1999. Seven-transmembrane proteins as odorant and chemosensory receptors. *Science* **286**:707–711.
- Nachman, M. W. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**:481–485.
- Nei, M., and W. H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**:5269–5273.
- Nickerson, D. A., S. L. Taylor, K. M. Weiss et al. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**:233–240.
- Payseur, B. A., and M. W. Nachman. 2000. Microsatellite variation and recombination rate in the human genome. *Genetics* **156**:1285–1298.
- Pilpel, Y., and D. Lancet. 1999. The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8**:969–977.
- Pritchard, J. K., and M. Przeworski. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **68**:1–14.
- Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160**:1179–1189.
- Przeworski, M., R. R. Hudson, and A. Di Rienzo. 2000. Adjusting the focus on human variation. *Trends Genet.* **16**:296–302.
- Reich, D. E., M. Cargill, S. Bolk et al. 2001. Linkage disequilibrium in the human genome. *Nature* **411**:199–204.
- Risch, N., K. Merikangas, S. A. Tishkoff et al. 1996. The future of genetic studies of complex human diseases. *Science* **273**:1516–1517.
- Rouquier, S., A. Blancher, and V. Giorgi. 2000. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc. Natl. Acad. Sci. USA* **97**:2870–2874.
- Sharon, D., G. Glusman, Y. Pilpel, M. Khen, F. Gruetzner, T. Haaf, and D. Lancet. 1999. Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. *Genomics* **61**:24–36.
- Slatkin, M., and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**:555–562.
- Stephan, W., H. E. W. Thomas, and M. W. Lenz. 1992. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theor. Popul. Biol.* **41**:237–254.
- Stephens, J. C., J. A. Schneider, D. A. Tanguay et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**:489–493.
- Tajima, F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics* **123**:597–601.
- . 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Tishkoff, S. A., R. Varkonyi, N. Cahinhinan et al. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**:455–462.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**:256–276.
- Whissell-Buechy, D., and J. E. Amooore. 1973. Letter: Odour-blindness to musk: simple recessive inheritance. *Nature* **245**:157–158.
- Wysocki, C. J., and G. K. Beauchamp. 1984. Ability to smell androstenone is genetically determined. *Proc. Natl. Acad. Sci. USA* **81**:4899–4902.
- Young, J. M., C. Friedman, E. M. Williams, J. A. Ross, L. Tonnes-Priddy, and B. J. Trask. 2002. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.* **11**:535–546.
- Younger, R. M., C. Amadou, G. Bethel et al. 2001. Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res.* **11**:519–530.
- Zhang, X., and S. Firestein. 2002. The olfactory receptor gene superfamily of the mouse. *Nat. Neurosci.* **5**:124–133.
- Zozulya, S., F. Echeverri, and T. Nguyen. 2001. The human olfactory receptor repertoire. *Genome Biol.* **2**(6):1–12.

Naruya Saitou, Associate Editor

Accepted September 16, 2002