Signatures of Selection in the Human Olfactory Receptor *OR5I1* Gene

Andrés Moreno-Estrada, Ferran Casals,* Anna Ramírez-Soriano,* Baldo Oliva,† Francesc Calafell,*‡ Jaume Bertranpetit,*‡ and Elena Bosch*‡*

*Unitat de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain; †Grup de Bioinformàtica Estructural (Unitat de Recerca en Informàtica Biomèdica-Institut Municipal d'Investigació Mèdica), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain; and ‡Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, Barcelona, Catalonia, Spain

The human olfactory receptor (OR) repertoire is reduced in comparison to other mammals and to other nonhuman primates. Nonetheless, this olfactory decline opens an opportunity for evolutionary innovation and improvement. In the present study, we focus on an OR gene, *OR5I1*, which had previously been shown to present an excess of amino acid replacement substitutions between humans and chimpanzees. We analyze the genetic variation in *OR5I1* in a large worldwide human panel and find an excess of derived alleles segregating at relatively high frequencies in all populations. Additional evidence for selection includes departures from neutrality in allele frequency spectra tests but no unusually extended haplotype structure. Moreover, molecular structural inference suggests that one of the nonsynonymous polymorphisms defining the presumably adaptive protein form of *OR5I1* may alter the functional binding properties of the OR. These results are compatible with positive selection having modeled the pattern of variation found in the *OR5I1* gene and with a relatively ancient, mild selective sweep predating the "Out of Africa" expansion of modern humans.

Introduction

Olfaction is one of the 5 sensory windows we have into the environment, and as such, genes encoding for molecules responsible for the interaction with odorants can be subject to strong selection (Issel-Tarver and Rine 1997). Humans are able to distinguish more than 10,000 different smells (Reed 1990), which are detected by sensory neurons through the interaction of olfactory receptors (ORs) with a variety of odorants in the lining of the nose. ORs are 7 transmembrane domain G protein–coupled receptors encoded by single-coding exon genes spanning just around 1 kb, which are usually arranged in gene clusters (Ben-Arie et al. 1994; Trask et al. 1998; Aloni et al. 2006). OR genes comprise the largest gene family in mammalian genomes (Gilad, Man, et al. 2003). Defined according to similarity in their amino acid sequence, several families (over 40% of identity) and subfamilies (identity >60%) are distinguished within the human OR superfamily (Glusman et al. 2001). Therefore, despite the general homology of ORs, their protein sequences are very diverse, reflecting the broad range of activating ligands. To the best of our knowledge, no experimentally determined structure of an OR protein exists but several potential odorant-binding residues along the transmembrane domains have been suggested (Hall et al. 2004; Man et al. 2004; Katada et al. 2005). Unfortunately, we are still far from understanding any clear relation between ligand properties and those of the OR protein (Malnic et al. 2004).

A substantial fraction of the human OR family members are pseudogenes carrying one or more frame disruptions (Glusman et al. 2001). Interestingly, the pseudogenization of OR genes has been described to occur more often in the human lineage than in any other primate lineage, probably reflecting a smaller human reliance on the sense of smell (Gilad, Man, et al. 2003). Besides the apparent relaxation of selective pressures, some OR genes have been found to remain intact in several primate species suggesting different levels of evolutionary constraints operating over the OR repertoire (Gilad, Man, et al. 2003). In addition, particular cases of positive selection in human OR genes have also been reported (Gilad et al. 2000; Gilad, Bustamante, et al. 2003; Nielsen et al. 2005). Gilad, Bustamante, et al. (2003) revealed patterns of variability consistent with positive selection acting on human intact OR genes in their study of diversity on 20 OR genes in humans and chimpanzee. Nielsen et al. (2005) described 4 genes involved in olfaction (*OR2W1*, *OR5I1*, *OR2B2*, and *C20orf185*) within the top 50 genes showing most evidence for positive selection from their pattern of synonymous and nonsynonymous substitutions after the comparison of 13,731 chimpanzee and human orthologs. Additional population data on 20 European Americans and 19 African Americans in the same study revealed a low divergence to polymorphism ratio for the *OR5I1* gene. This result led the authors to invoke the possibility that balancing selection may have acted on the ORs.

Here we present data on single nucleotide polymorphism (SNP) allele and haplotype frequency variation for the *OR5I1* gene region in 39 human populations representing all major regions of the world. Two nonsynonymous and several synonymous polymorphisms characterize the commonest *OR5I1* haplotype from the more ancestral-like form of the receptor. We found signatures of natural selection and attempted to model the structure of the different proteins forms to explore the functional relevance of the probably adaptive amino acid changes.

Materials and Methods

Samples

We analyzed the Human Genome Diversity Cell Line Panel-Centre d'Etude du Polymorphisme Humain, which contains 1,064 DNA samples from individuals representing 51 populations globally distributed (Cann et al. 2002). For further analyses, we used the H971 subset recommended by Rosenberg (2006), in which atypical, duplicated individuals and deduced first-degree relatives have been removed. Samples in which genotypes for at least 50% of the SNPs

Key words: olfactory receptors, single nucleotide polymorphisms, human variation, positive selection.

E-mail: elena.bosch@upf.edu.

Mol. Biol. Evol. 25(1):144–154. 2008
doi:10.1093/molbev/msm240
Advance Access publication November 2, 2007

Table 1
Summary Statistics for *OR5I1* Variation in the 39 Worldwide Populations

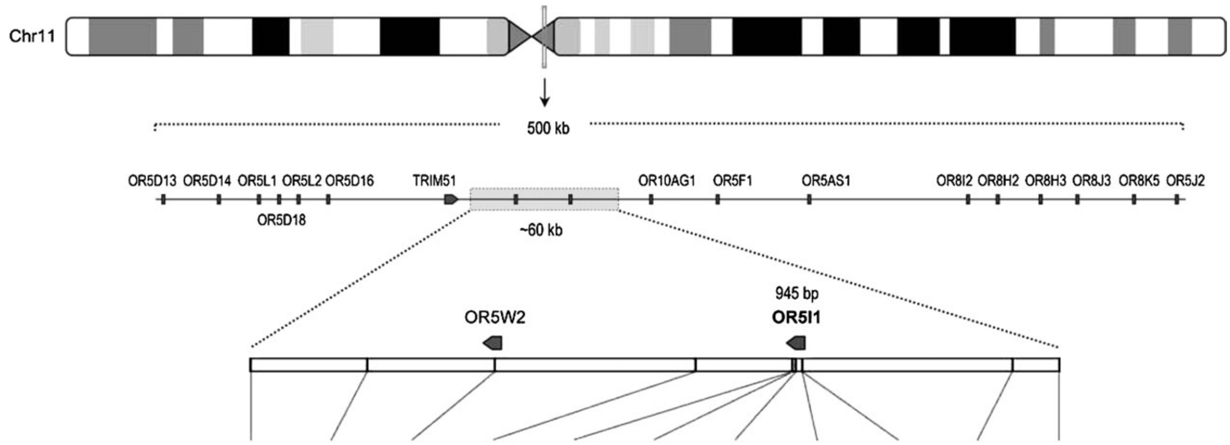
Population	2*N*[a]	S[b]	MAF <0.10[c]	MAF >0.40[d]	DAF <0.20[e]	DAF >0.80[f]	K[g]	H[h]
Sub-Saharan Africa								
Bantu	36	4	6	0	3	4	5	0.708 ± 0.002
Biaka Pygmies	54	9	3	0	3	2	7	0.793 ± 0.001
Mbuti Pygmies	26	9	7	2	3	3	7	0.742 ± 0.003
Mandenka	46	7	7	0	3	3	6	0.770 ± 0.001
San	12	10	7	2	3	3	6	0.878 ± 0.004
Yoruba	44	7	6	0	3	3	5	0.735 ± 0.001
Middle East–North Africa								
Mozabite	58	10	6	0	4	4	6	0.599 ± 0.004
Palestinian	100	10	6	0	4	4	7	0.610 ± 0.002
Bedouin	94	10	8	0	4	5	6	0.487 ± 0.003
Druze	88	10	7	0	4	5	5	0.468 ± 0.004
Europe								
French	56	10	6	0	4	5	4	0.462 ± 0.005
Basque	48	10	10	0	4	5	4	0.301 ± 0.007
Orcadian	28	10	1	0	2	3	4	0.595 ± 0.008
Sardinian	56	10	2	0	2	0	5	0.583 ± 0.004
North Italian	42	9	2	0	4	5	3	0.296 ± 0.007
Adygei	34	10	2	0	4	5	5	0.449 ± 0.010
Russian	50	10	9	0	4	5	4	0.320 ± 0.007
Central–South Asia								
Balochi	48	10	7	0	4	5	5	0.335 ± 0.007
Brahui	50	9	2	0	4	5	5	0.322 ± 0.007
Burusho	50	10	10	0	4	5	4	0.154 ± 0.005
Hazara	46	9	8	0	4	5	4	0.241 ± 0.006
Kalash	48	3	10	0	4	5	2	0.082 ± 0.003
Makrani	50	10	8	0	4	5	6	0.389 ± 0.007
Pathan	48	10	8	0	4	5	4	0.301 ± 0.007
Sindhi	48	10	7	0	4	5	5	0.426 ± 0.007
Northwest China	58	9	7	0	4	5	4	0.276 ± 0.005
East Asia								
Northeast China	76	3	7	0	4	5	2	0.212 ± 0.003
South China	132	3	7	0	4	5	3	0.216 ± 0.002
Han	88	4	10	0	4	5	4	0.171 ± 0.003
Yakut	50	10	8	0	4	5	4	0.257 ± 0.006
Cambodian	20	3	7	0	4	5	2	0.268 ± 0.013
Japanese	58	3	10	0	4	5	2	0.100 ± 0.003
Oceania								
Nasioi	26	1	9	0	4	5	2	0.271 ± 0.010
Papuan	34	3	7	0	4	5	2	0.300 ± 0.007
America								
Pima	28	0	10	0	4	5	1	0.000 ± 0.000
Maya	44	1	10	0	4	5	2	0.046 ± 0.002
Colombian	14	0	10	0	4	5	1	0.000 ± 0.000
Karitiana	28	0	10	0	4	5	1	0.000 ± 0.000
Surui	18	0	10	0	4	5	1	0.000 ± 0.000
Worldwide	1934							

[a] Number of chromosomes.
[b] Number of segregating sites.
[c] Number of SNPs with MAF lower than 10%. Includes SNPs fixed in one population but polymorphic elsewhere.
[d] Number of SNPs with MAF greater than 40%.
[e] Number of SNPs with DAF lower than 20%.
[f] Number of SNPs with DAF greater than 80%.
[g] Total number of haplotypes.
[h] Haplotype diversity.

failed were also excluded. In order to maximize sample sizes, population samples were regrouped into 39 populations based on geographic and ethnic criteria as in Gardner et al. (2006). For part of the analysis, populations were further grouped into 7 geographical regions (see table 1).

Single Nucleotide Polymorphisms

We have genotyped a total of 11 SNPs in the OR5I1 gene region (fig. 1): 2 synonymous and 3 nonsynonymous SNPs in the 945 bp OR5I1 coding region plus 6 additional SNPs covering both 5′ and 3′ flanking regions up to around 30 kb. The coding SNPs genotyped were all those available in the single nucleotide polymorphism database (dbSNP) at the start of the study, which in turn passed the multiplexing design of SNPlex. Ancestral alleles were those recovered from the chimpanzee and/or the macaque genome sequences (panTro2, March 2006 assembly and rheMac2, January 2006 assembly, respectively) except for one SNP (rs3923162), where none of the human alleles (T/C) were found in the chimpanzee sequence (which carries a G) and for which no other mammal ortholog sequence



	rs9943659	rs7940445	rs2457239	rs7115131	rs9665861	rs9665863	rs4465383	rs9666086	rs4367963	rs3923162	rs12577044
SNP alleles [a]	C/T	T/C	T/C	C/T	C/T	C/A	A/G	G/A	A/G	T/C	C/A
SNP type [b]	FLK3	FLK3	FLK3	FLK3	CNS	CS	CS	CNS	CNS	FLK5	FLK5
Protein residue	---	---	Cys/Arg	---	Val/Ile	Pro/Pro	Ile/Ile	Ser/Phe	Leu/Ser	---	---
Inter marker distance (bp)	---	8,491	9,193	14,491	6,976	49	237	403	78	15,168	3,329
Mean DAF [c]	0.96	0.08	0.00	0.96	0.05	0.04	0.85	0.95	0.08	0.95	0.85
F_{ST} value [d]	0.05	0.09	---	0.05	0.05	0.06	0.07	0.05	0.03	0.05	0.07

[a] The allele named first corresponds to the ancestral state for each SNP (see Materials and Methods).

[b] FLK3 stands for 3' flanking, FLK5 stands for 5' flanking, CNS stands for coding non synonymous,and CS stands for coding synonymous.

[c] Derived allele frequency averaged across 39 populations.

[d] F_{ST} parameter calculated across 39 populations.

FIG. 1.—Schematic view of the *OR5I1* region and information on the genotyped SNPs. *OR5I1* lies in the centromeric region of the chromosome 11, which contains a large proportion of ORs belonging mainly to the OR5 family as detailed in the enlarged 500-kb region below the ideogram. The dashed line box in gray denotes the genotyped area spanning around 60 kb, within which the relative position of the SNPs genotyped is shown. SNP type categories are named with respect to the *OR5I1* gene, but amino acid changes are shown for all coding SNPs in the region.

was available. SNPs were typed using the SNPlex Genotyping System from Applied Biosystems within a larger set of 365 SNPs covering 18 additional genes (see below) and following the manufacturer's standard protocol. Allele separation was performed on an Applied Biosystems 3730 analyzer and besides the automated allele calling and quality metrics provided by GeneMapper Software 3.5, allele calling was always reviewed manually.

Reference Gene Set

Seventeen genes belonging to different functional categories with evidence of acceleration in the human lineage were chosen from Clark et al. (2003): *AIRE*, *ALDH6A1*, *BCKDHA*, *CLDN8*, *CST2*, *DACT1*, *DIAPH1*, *EYA4*, *FOXI1*, *GIP*, *GSTZ1*, *HCLS1*, *IL15RA*, *IL1RL2*, *MRPL35*, *TECTA*, and *TMPRSS2*. These 17 gene loci together with the *LCT* gene have been used as reference data in both minor allele frequencies (MAFs) and derived allele frequencies (DAFs) threshold analyses. SNPs covering such genes were selected with a marker density of 1 SNP every 5–10 kb inside each gene and adding several extra SNPs up to 30 kb in both flanking regions. Preference was given to SNPs with a MAF over 10%, which were compiled from HapMap (Release 7 May 2004) and dbSNP (Build 121 June 2004) databases. Additionally, most coding SNPs and other functional SNPs identified using PupaSNP Finder (Conde et al. 2004) were also included for analysis regardless of their allele frequency or validation status. Note that SNP selection in the *OR5I1* region fulfilled the same criteria. Genotyped SNP density for the whole reference gene set was 0.17 and that of *OR5I1* was slightly above the average (0.19), but well within the distribution: 5 genes in the reference set had higher SNP densities than *OR5I1*. Ancestral alleles for these SNPs were deduced as indicated above for *OR5I1*.

Basic Analysis

Genotype data were collected and stored in a database within the SNPator web environment (http://bioinformatica.cegen.upf.es), where part of the analyses such as control for replicate samples along sample plates and basic analysis such as allele frequencies, expected heterozygosity and Hardy–Weinberg equilibrium tests were performed. Arlequin (Schneider 2000) was used to calculate F_{ST} values between the 39 populations studied with a locus-by-locus analysis of molecular variance (Excoffier et al. 1992).

Haplotype Analysis

Haplotypes were inferred from unphased genotype data using the Bayesian statistical method in program PHASE 2.1 (Stephens et al. 2001) using the default parameter set with 1,000 iterations. Relationships between inferred nonrecombinant haplotypes were investigated using the median-joining network algorithm (Bandelt et al. 1995, 1999) within Network 4.201 software. In order to facilitate visualization of phylogenetic relationships, variation at the recurrent rs12577044 was not considered. Human reference sequence for *OR5I1* and its ortholog sequence in chimpanzee were extracted from Ensembl (Gene IDs ENSG00000167825 and ENSPTRG00000003615). Ortholog sequences were translated and aligned with ClustalW (Thompson et al. 1994).

Analysis of Signatures of Selection

Departures from neutral evolution were tested by means of 4 different methods: 1) comparison of the SNP allele frequency spectrum to a reference set (considering the frequency of either the least frequent or the derived allele), 2) the ratio of synonymous to nonsynonymous substitutions, 3) classical neutrality tests such as Tajima's *D* and others in the resequencing data by Nielsen et al. (2005), and 4) the extent of haplotype homozygosity. Further details on how these methods were carried out can be found as supplementary note S1 (Supplementary Material online).

Three-Dimensional Structural Prediction Model for *OR5I1*

Protein sequences for the human *OR5I1* major variant and its orthologs in the following species were retrieved from either the Ensembl or the Human Olfactory Receptor Data Exploratorium database (Safran et al. 2003): human (Swissprot ID: Q13606), chimpanzee (CONTIG970.38-26808), macaque (ENSMMUG00000016153), rat (NM_001000329), mouse (NM_146646), dog (c*OR5I1*), and opossum (Modo-*OR5I1*). Ancestral-like and major-derived *OR5I1* variants were inferred on the major protein sequence from resequencing data. Multiple alignment of the aforementioned protein sequences and those of *OR5U1* (Swissprot ID: Q9UGF5) and bovine rhodopsine (Swissprot ID: P02699) was performed using Hidden Markov Models and considering Pfam domains (Finn et al. 2006). Three-dimensional structural models of chimpanzee and the 3 human *OR5I1* variants were obtained using the 3-dimensional crystalline structure of bovine rhodopsin (PDB ID1F88A) as template and Modeller (Sali and Blundell 1993). Particular codon changes between *OR5I1* variants were explored and visually analyzed with RasMol 2.7.1.1. (Bernstein 2000) and Prepi (http://www.sbg.bio.ic.ac.uk/prepi).

Results

Allele Frequencies and Haplotype Phylogenetic Reconstruction

Details for the *OR5I1* region and the 11 SNPs analyzed are shown in figure 1. All SNPs and populations were in Hardy–Weinberg equilibrium after Bonferroni correction for multiple tests. With the exception of sub-Saharan Africa, allele frequencies were rather homogenous across continental regions (fig. 2). Around half of the SNPs analyzed appeared to have derived frequencies equal or greater than 85% across all populations, whereas rs2457239 was found to be fixed. F_{ST} values between the 39 populations analyzed were rather low (fig. 1), with an average across SNPs of around 0.06.

Following haplotype reconstruction, we identified a total of 16 different haplotypes across the 39 populations analyzed (fig. 3). A predominant haplotype (H10) was found representing over 50% of the chromosomes in every population except for the sub-Saharan Africans, where other haplotypes were found at common frequencies (supplementary fig. S1 and table S1, Supplementary Material online). Indeed, the haplotype diversities in sub-Saharans were significantly higher than those in any non sub-Saharan population (table 1). Considering the recombination rates of the *OR5I1* region estimated from linkage disequilibrium (Myers et al. 2005) and comparing the human haplotypes with the chimpanzee ancestral positions, we suggest a minimum of 3 recombinant haplotypes and several recurrent substitutions (fig. 3 and supplementary note S2 [Supplementary Material online] for recombinant and recurrence inference). Haplotype phylogenetic relationships were explored by means of a median-joining network (fig. 4) except for those haplotypes inferred as recombinants. The 3 commonest human haplotypes (H10, H12, and H16) were found worldwide and clustered quite together in the network, being 1-step or 2-step neighbors but far away from the less frequent ancestral-like haplotypes.

MAF and DAF Threshold Analyses

The proportion of SNPs in the *OR5I1* region having a MAF above 0.4 and below 0.1 was counted for each population and compared with those equivalent proportions obtained for 18 other genes studied in the same 39 populations. In comparison to our reference gene set, *OR5I1* displayed a relatively constant pattern of excess of low-frequency minor alleles across most populations (results not shown). Most of the populations presented more

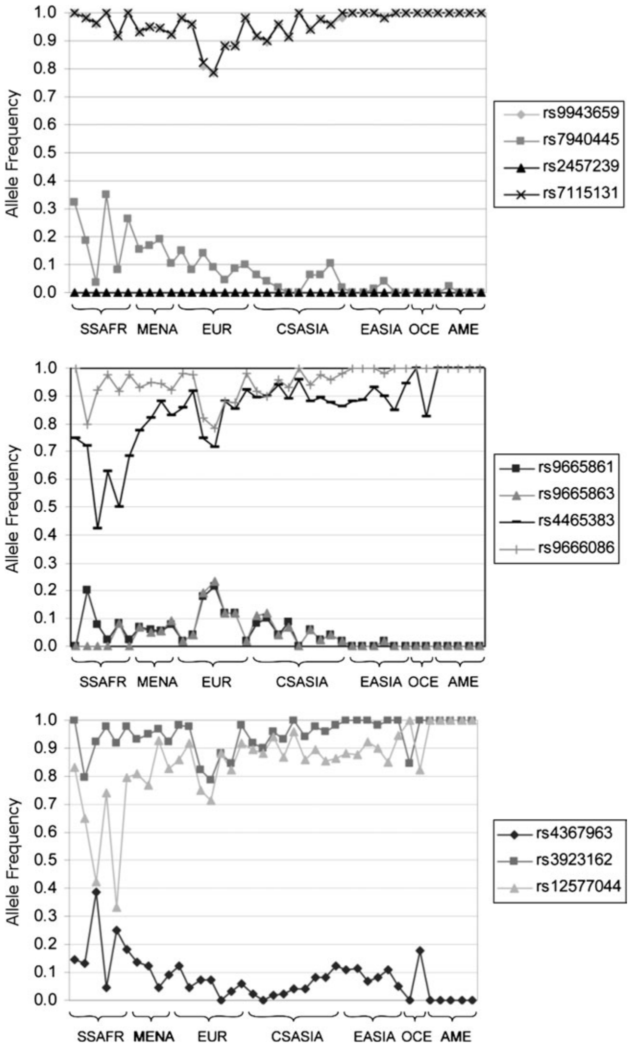


FIG. 2.—DAFs for the 11 SNPs genotyped across 39 different populations. Every point represents one population, displayed from left to right as in table 1. Curly brackets on the *x* axis indicate groups of populations belonging to the following continental regions: Sub-Saharan Africa (SSAFR), Middle East–North Africa (MENA), Europe (EUR), Central–South Asia (CSASIA), East Asia (EASIA), Oceania (OCE), and America (AME). The C allele is plotted for rs39231162, but no ancestrality inference was possible from the chimpanzee genome or any other close-related ortholog sequences.

	rs9943659	rs7940445	rs7115131	rs9665861	rs9665863	rs4465383	rs9666086	rs4367963	rs3923162	rs12577044	
Chimp	C	T	C	C	C	A	G	A	G	C	
H1	.	.	.	T	.	.	.	.	T	.	2
H2	.	.	.	T	A	.	.	.	T	.	70
H3	.	.	T	T	A	.	.	.	T	.	1
H4	.	.	.	T	A	.	.	.	T	A	1
H5	.	.	.	T	A	.	.	.	C	A	2
H6	.	.	.	T	A	G	.	.	T	A	1
H7	T	.	T	T	.	G	.	.	T	.	14
H8	T	.	T	.	.	G	A	.	T	.	1
H9	T	.	T	.	.	G	A	.	T	A	4
H10	T	.	T	.	.	G	A	.	C	A	1464
H11	T	.	T	.	.	G	A	.	C	.	15
H12	T	C	T	.	.	G	A	.	C	A	157
H13	T	.	T	.	A	G	A	.	C	A	4
H14	T	.	T	.	.	.	A	.	C	A	28
H15	T	.	T	.	.	.	A	.	C	.	16
H16	T	.	T	.	.	.	A	G	C	.	154

FIG. 3.—Graphical representation of human *OR5I1* haplotypes. Each polymorphic variant is displayed below the corresponding chimpanzee ancestral position. Ancestral chimpanzee-like alleles are indicated with dots, whereas underlined positions denote possible recurrent and back mutations. Different shades indicate different amino acid sequences: light gray, ancestral like (Leu-Ser-Ile); medium gray, major (Leu-Phe-Val); and dark gray, major derived (Ser-Phe-Val). Sequences with alternating shades indicate putative recombinants.

than 5 SNPs on the *OR5I1* region with MAFs <0.1 and only the Mbuti Pygmies and the San population had SNPs with a MAF >0.4 (see table 1). The proportion of SNPs at high DAFs (>0.8) against low DAF (<0.2) for each population and gene analyzed is plotted in figure 5*A*. *OR5I1* clearly stands out as an outlier in 32 out of 39 populations, indicating an excess of high-frequency derived alleles in this gene region across all major geographical regions of the globe. Moreover, the non-*OR5I1* data points that lie out of the empirical confidence interval in figure 5*A* corresponded all to Amerindians or Oceanians, which may reflect the effects of genetic drift and small sample sizes rather than any differential selective signature.

In order to test whether any particular ascertainment bias in our set of genes could affect our results and to be able to provide statistical support to our observations, we compared our data for French, Han Chinese, and Yorubas with 2 data sets produced by Walsh et al. (2006): 1) a set of SNPs in 64 gene regions related to immune function genotyped in Europeans from Utah, Han Chinese from Guanxi, and Yoruba Nigerians and 2) simulated data points according to a demographic model specific for each of these 3 continental groups. On both data sets, SNPs were selected with criteria equivalent to ours. As shown in figure 5*B*, the proportion of SNPs with low MAFs in Yorubas and Han Chinese is particularly high in *OR5I1* as compared with both data sets. Taking the 1,000 simulated distributions, the probability that *OR5I1* behaves as a neutral gene in these 2 populations is $P < 0.03$. More importantly, and for all 3 populations, we found that *OR5I1* shows one of the highest proportions of SNPs with DAF >0.8 ($P \leq 0.002$).

OR5I1 Nonsynonymous Polymorphic Variation

It has been recognized that several amino acids in the transmembrane domains of the ORs form a ligand-binding pocket (Man et al. 2004; Katada et al. 2005). Moreover, experimental data suggest that the ligand specificity of an OR can be manipulated by point mutations in such domains, allowing any OR to change its affinity for certain odorant molecules (Katada et al. 2005). The 3 *OR5I1* nonsynonymous SNPs genotyped fall in different transmembrane domains of the receptor and were found to define

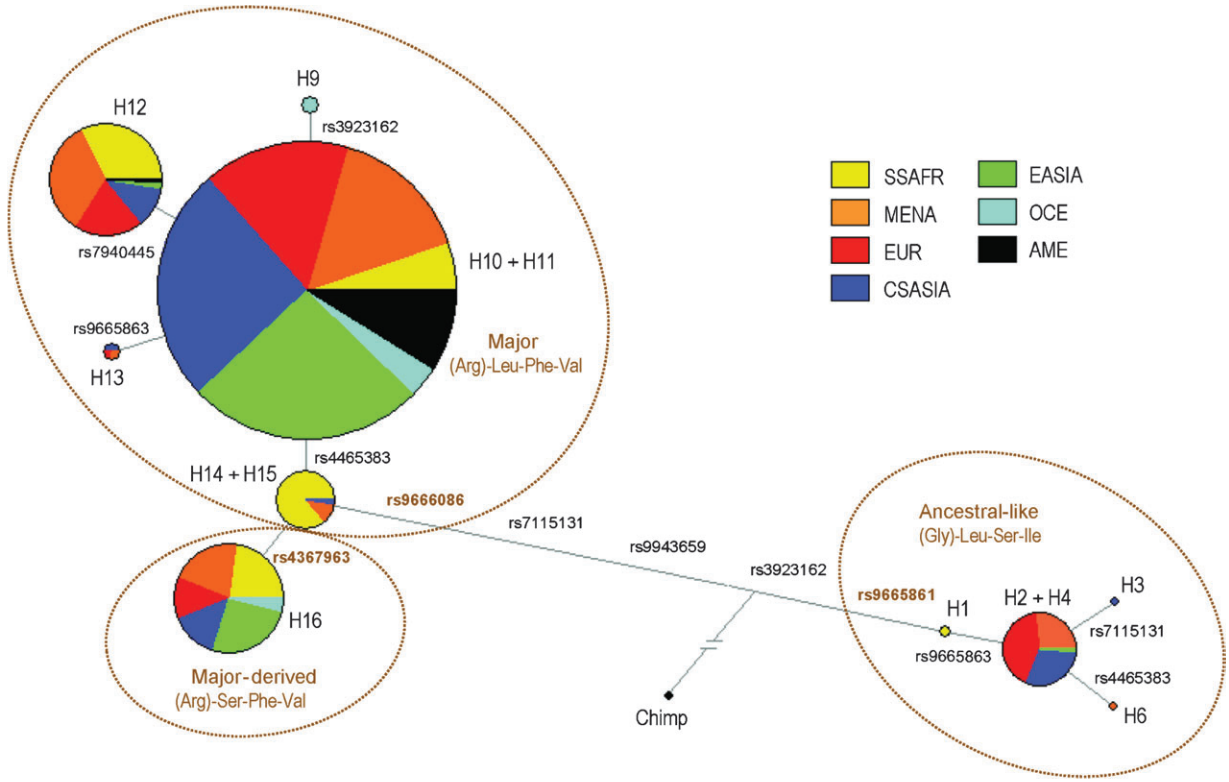


FIG. 4.—Median-joining network of OR5I1 haplotypes. Nodes in the median-joining network are proportional to frequencies and branch lengths to the number of polymorphic base substitutions, indicating nonsynonymous changes in brown. For the branch tracing back to the chimpanzee, 6 divergent positions must be taken into account besides rs3923162, where none of the human alleles (C/T) were found in the corresponding chimpanzee OR5I1 flanking sequence (which carries a G). Variation at the recurrent rs12577044 was not considered. Continental regions are denoted as in figure 2. Haplotypes coding for the same amino acid sequence are enclosed in brown ovals.

3 protein variants for *OR5I1* in our worldwide human panel (see fig. 4), which we will refer as ancestral like (Leu-Ser-Ile), major (Leu-Phe-Val), and major derived (Ser-Phe-Val). There are 2 *OR5I1* nonsynonymous substitutions publicly available but not typed in this study: rs17597625 (G6R) and rs11607055 (L307I). HapMap and resequencing data in 19 African Americans and 20 European Americans by Nielsen et al. (2005) show rs11607055 to be monomorphic for leucine and the derived allele of rs17597625 in complete linkage disequilibrium with the derived allele of rs9666086 (F76S). Human resequencing data for *OR5I1* revealed 3 additional nonsynonymous variations, which were singletons occurring on the major form of the protein (Nielsen et al. 2005).

Physicochemical distances of Grantham (1974) for the amino acid replacements corresponding to the 4 segregating nonsynonymous *OR5I1* SNPs as well as their functional effect predicted by Polyphen (Ramensky et al. 2002) indicate that only those replacements observed at codon positions 50 and 76 could potentially affect the corresponding α-helices on the secondary structure of the *OR5I1* receptor (supplementary table S2, Supplementary Material online). Man et al. (2004) predicted several putative odorant-binding site residues on the OR proteins using human–mouse comparisons. Although none of the identified nonsynonymous *OR5I1* variants in humans correspond to these putative odorant-binding sites, rs9666086 (F76S) lies just next to the only predicted binding site position on the second transmembrane region. This seems relevant because the change of serine to phenylalanine in codon 76 characterizes the major form of the *OR5I1* gene in humans. It should also be noticed that functionally relevant species-specific novelties as the ones evaluated here could not be detected by the ortholog–paralog comparison approach of Man et al. (2004).

We further evaluated the amino acid residue changes among the different *OR5I1* protein forms by means of structural modeling, using the bovine rhodopsin, the only structural template available for G protein–coupled receptors. As expected, none of the amino acid replacements at codon positions 6 and 306 appear to destabilize the modeled structure. Although the Leu to Ser change at codon position 50 at the end of the first transmembrane domain involves the introduction of a polar amino acid in a hydrophobic environment, such replacement does not appear to have significant structural impact on the receptor (results not shown). However, the Ser to Phe replacement at codon position 76 on the second transmembrane domain might imply a significant structural change on the model. The substitution of a polar amino acid by an aromatic one is very likely to provide higher overall stability to the structure in this particular hydrophobic environment of the membrane. But more important than that, in this case, it seems to force the movement of a tyrosine on the adjacent transmembrane domain (codon position 37) modifying the spatial configuration and the interaction among the first and the second transmembrane domains (see fig. 6). We can therefore conclude that one of the

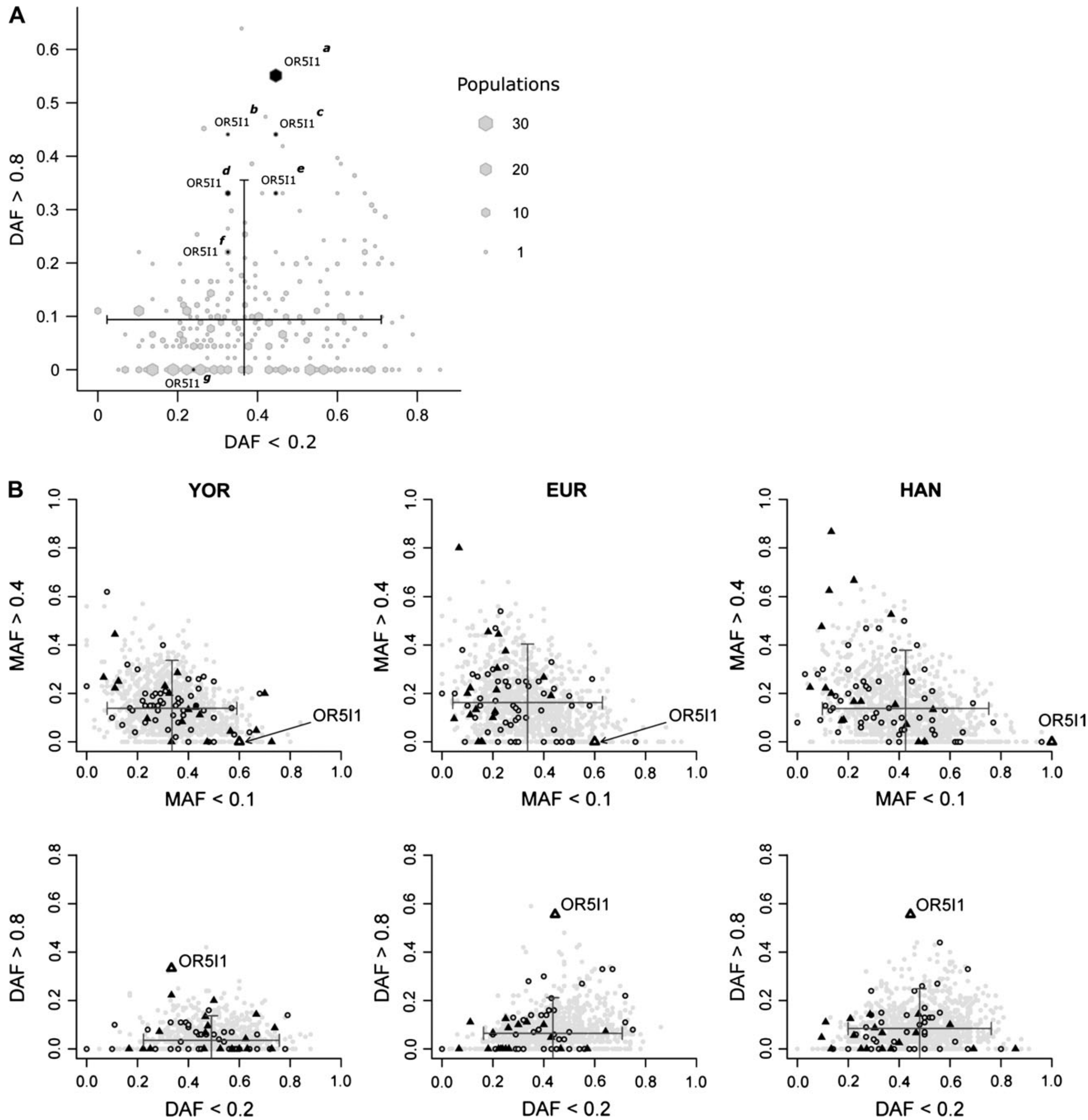


FIG. 5.—Allele frequency threshold analyses for signals of selection. (*A*) The proportion of SNPs with DAF above 80% and DAF below 20% is plotted in a dispersion diagram for every gene-population pair analyzed in this study. In total 741 data points are displayed; however, a single dot in the plot may represent several overlapping results. The area of hexagonal dots is proportional to the number of gene-population instances observed. Vertical and horizontal lines on the *y* and *x* axis represent 95% confidence intervals. All points involving *OR5I1* are shown as black hexagons and reference genes in gray. *a* corresponds to Bedouin, Druze, French, Basque, North Italian, Adygei, Russian, Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, Sindhi, Northwest China, Northeast China, South China, Han, Yakut, Cambodian, Japanese, Nasioi, Papuan, Pima, Maya, Colombian, Karitiana, and Surui; *b* represents Bantu; *c*, Palestinian and Mozabite; *d*, Mbuti Pygmies, Mandenka, Yoruba and San; *e*, Orcadian; *f*, Biaka Pygmies; and *g*, Sardinian. (*B*) Allele frequency analysis for signals of selection based on MAFs and DAFs in Yorubas, Europeans, and Han Chinese, respectively. In each plot, the corresponding data point of the *OR5I1* gene is indicated by an open triangle. As inference for rs3923162 allele ancestrality was ambiguous, we did not include this SNP in the analysis. Full triangles indicate the 18 comparative reference genes genotyped in this study; open circles, experimental data on gene regions related to immune function from Walsh et al. (2006) and gray full circles, simulation data from Walsh et al. (2006). Vertical and horizontal lines on the *y* and *x* axis represent 95% confidence intervals based on simulated data.

amino acid replacements that characterizes the major form of the protein definitively affects the molecular environment of the odorant-binding pocket of the *OR5I1* receptor and thus it could be selectively relevant.

Five out of the 6 diverging sites between human and chimpanzee sequences are nonsynonymous (fig. 4). Sequence comparison with 5 different mammal orthologs shows only one of these diverging substitutions happening

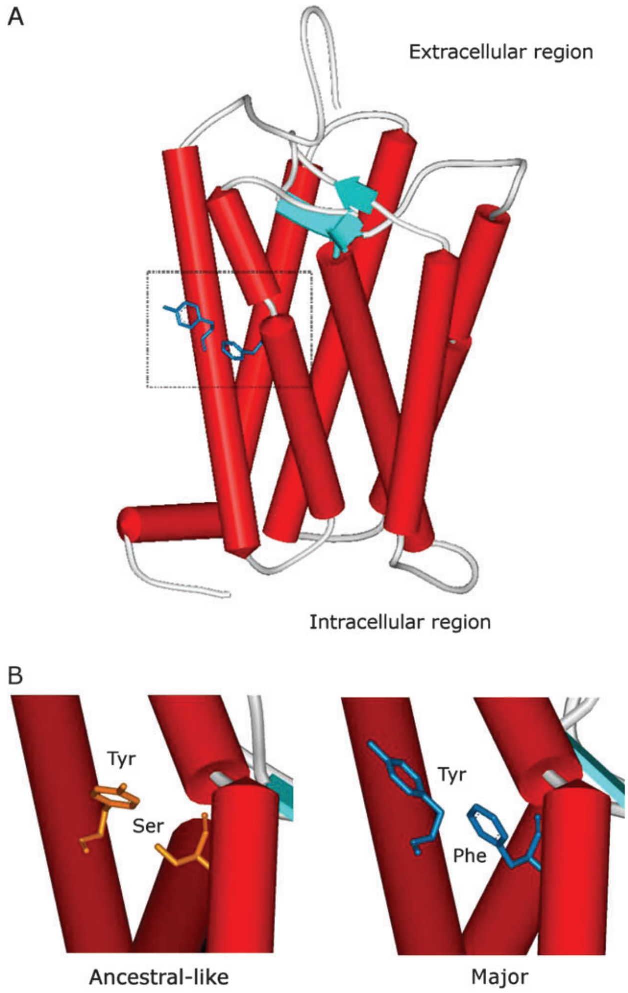


FIG. 6.—Three-dimensional model of the *OR5I1* receptor. (*A*) General view of the major form (Leu-Phe-Val) of the receptor. (*B*) Close-up of amino acid position 76 comparing the major (Leu-Phe-Val) and the ancestral-like (Leu-Ser-Ile) form of the *OR5I1* protein. Note the alternate position of Tyr37 as consequence of the Phe76Ser replacement.

in the human branch: a Ser to Phe replacement that lies in the third intracellular loop and that it is predicted to be benign by PolyPhen.

Signatures of Positive Selection

OR5I1 was found within the top 50 genes showing most evidence for positive selection from their pattern of synonymous and nonsynonymous substitutions after the comparison of 13,731 chimpanzee and human orthologs (Nielsen et al. 2005). However, a likelihood ratio test failed to reject the null hypothesis of $d_N/d_S = 1$ versus the alternative hypothesis of $d_N/d_S > 1$ for *OR5I1* in the same study. Here, a branch site likelihood method (Zhang et al. 2005) was not powerful enough to detect accelerated protein evolution in any human branch of a phylogeny containing the 3 human *OR5I1* protein variants and its orthologues in chimpanzee, macaque, mouse, rat, and opossum. Accelerated protein evolution was not detected on the chimpanzee branch either. But notably, the human branch leading to the major form of the *OR5I1* did show a unique significant posterior Bayesian probability of 97.9% of being a target of selection precisely for codon site 76.

We further investigated the genetic footprint of selection from the resequencing data available in 20 European Americans and 19 African Americans in *OR5I1* (Nielsen et al. 2005). As shown in table 2, all the neutrality statistics were negative in both populations, a pattern consistent with positive selection, a population expansion, or purifying selection on slightly deleterious alleles. In order to evaluate the significance of these values and, at the same time, to discard the possible effect of demographic factors such as population expansions, we have simulated neutral distributions including the corresponding inferred human demographic history (Schaffner et al. 2005). Tajima's *D* and Fu and Li's *F* and *F** have resulted significant in European Americans but only Fay and Wu's *H* is significant in the case of the African Americans (table 2). These results seem to indicate that positive selection may have been more intense in Europeans than in Africans. We also produced coalescent neutral simulations with recombination but with a constant population size; in those, Fu and Li's *D* and *D** became significant in European Americans (with *P* values of 0.024 and 0.036, respectively) and Fay and Wu's *H* was no longer significant in African Americans ($P = 0.053$). Tajima's *D* was also contrasted against the empirical distribution for 293 genes resequenced in European Americans and African Americans (Seattle SNPS, http://pga.gs.washington.edu). In an empirical 1-tailed test, Tajima's *D* was significant in European Americans ($P = 0.020$), but it was not in African Americans ($P = 0.070$), as in the original demographic simulations. In summary, the conclusions about the significance of the neutrality tests seemed to be robust to the demographic model used to provide statistical significance.

A well-recognized signature of recent positive selection is the finding of long-range haplotypes surrounding those selected alleles that have risen in frequency rapidly enough for avoiding recombination to have substantially

Table 2
Selection Statistics for the *OR5I1* Gene

Population	$2N^a$	Tajima's *D*	Fu and Li *D*	Fu and Li *F*	Fu and Li *D**	Fu and Li *F**	Fay and Wu *H*
African American	36	−1.376	−1.592	−1.816	−1.426	−1.655	−3.241
P value*		n.s.	n.s.	n.s.	n.s.	n.s.	0.018
European American	36	−1.752	−2.220	−2.444	−2.101	−2.323	−1.565
P value*		0.021	n.s.	0.044	ns	0.042	n.s.

[a] Number of chromosomes with enough sequencing information available (Nielsen et al. 2005).
*P value < 0.05.

broken down their allele association with alleles at nearby loci. Extended haplotype analysis of our data was not feasible due to the low recombination rate (Myers et al. 2005) of the 60-kb *OR5I1* genotyped region. Consequently, we explored HapMap data (Release 21 July 06) extending 650-kb centromere proximal and 4.3-Mb centromer distal to the *OR5I1* genotyped region for evidence of recent selection using both the extended haplotype homozigosity and the relative extended haplotype homozigosity (REHH). In none of the 3 HapMap population groups, any core haplotype involving the *OR5I1* gene region analyzed here was found to represent a significant outlier relative to other core haplotypes within the considered extended 5-Mb region in chromosome 11 (results not shown). Similarly, the query for evidence of selection in any HapMap SNP around the *OR5I1* region by means of the integrated haplotype score (iHS) (Voight et al. 2006) demonstrated no significant signal of recent selection.

Discussion

Several lines of evidence suggest that positive selection has modeled the pattern of variation found in the *OR5I1* gene: 1) a larger than expected proportion of SNPs show derived alleles at high frequencies, 2) allele frequency spectra in full sequence data show an excess of low-frequency alleles in European Americans, and 3) a nonsynonymous replacement at codon 76 is likely to have structural repercussions in the odorant-binding site of the *OR5I1* protein and thus be functionally relevant (and, as discussed below, adaptive).

One could argue that selection may have operated elsewhere in the vicinity of this gene or that random genetic drift in early human populations could have driven a particular amino acid replacement nearly to fixation. However, the demonstration that rs9666086 (F76S) has structural consequences on the putative odorant-binding region of the *OR5I1* receptor together with the finding of a pattern of genetic variation around this gene region compatible with the action of selection render less plausible any of these 2 alternative scenarios. Both the acquisition of a new ligand-binding capability or the modification of a particular odorant perception could improve the overall degenerated human OR gene repertoire. Thus, it is feasible that such changes could have been selected for and spread worldwide. The low F_{ST} values observed across the 39 worldwide populations analyzed seem to point to a homogenous selective pressure across the human species. This may suggest that the major form of the *OR5I1* protein links to an environmentally ubiquitous odorant or that it could detect a human-specific olfactory signal. It is also significant that the predominant form of *OR5I1* is found at high frequencies in all major geographical regions but has not reached fixation. Such frequency and distribution lead us to suggest a date of origin for the appearance of the major form of the *OR5I1* protein previous to the "Out of Africa" dispersion of modern humans but not a much older age. Although signatures of selection such as long haplotypes around high frequency recently selected alleles are believed to persist only around 30,000 years, high-frequency derived alleles reveal selective events over the last 80,000 years approximately (Sabeti et al. 2006). Therefore, the absence of signals of recent selection in any HapMap population as explored by means of the REHH or by the iHS statistics together with the finding of high-frequency derived alleles in all human populations seem to indicate that *OR5I1* may have been a target of selection already in early human evolution before the dispersal of modern humans. The clearest signal of selection that we have detected on the pattern of *OR5I1* human variation is an excess of derived alleles at high frequency, whereas the excess of rare alleles is marginally significant both according to the MAF threshold analysis and to most of selection statistics obtained from resequencing data. However, we believe these rare *OR5I1* alleles are not, as usually interpreted, new variants appearing after a selected allele has swept to fixation; they probably are the alleles linked to a variant or variants that are in the process of being outcompeted by the functional variant that we suggest is adaptative. That situation can be interpreted as a partial selective sweep; because derived alleles are present in all continents, this selective sweep may have been in process for tens of thousands of years and thus may involve very low selection coefficients. The different proportions of the major haplotype clusters in Africans and non-Africans may be the result of different local selective pressures or may also reflect the Out of Africa bottleneck (Mateu et al. 2002). Note though that such differences have not resulted in particularly high F_{ST} values. *OR5I1* has a high polymorphism to divergence ratio in humans in comparison to chimpanzees, a pattern that has been suggested to possibly result from balancing selection (Nielsen et al. 2005). However, neither the global allele frequency distributions observed nor the MAF threshold analysis agree with such hypothesis.

Olfaction is one of the most ancient senses, with which mammals monitor the environment allowing the recognition of mates, offspring, predators, tainted food, and chemical dangers. OR genes have experienced an important loss in humans, probably reflecting a lesser reliance on the sense of smell in comparison to other primates. Although the accumulation of mutations on any redundant gene copy of the genome is more likely to involve gene silencing than to eventually evolve into a new function, such situation could be facilitated in a scenario of general relaxation of selective constraints. Within a general scenario of ongoing pseudogenization and loss of selective constraints for the G protein–coupled bitter-taste receptor repertoire in higher primates, signatures of positive selection have also been detected for the human bitter-taste receptor TAS2R16 (Wooding et al. 2004; Soranzo et al. 2005). Even though humans as a species show reduced perception abilities in comparison to other mammals and other primates, the innovation and subsequent conservation of specific sensory functions may have spread, as seen in the bitter taste gene. Here we show that *OR5I1* may be an example of an OR gene running counter to the general trend toward loss of olfactive function in humans and actually providing what might be a clearly adaptive new role. Alas, given the actual state of the knowledge about ORs and their ligands, such a new and specific role remains, for the time being, elusive.

Supplementary Material

Supplementary notes S1 and S2, figure S1, and tables S1 and S2, are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We want to acknowledge and thank K. Kristjansdottir, M. Sikora, M. Melé, T. Marquès, and A. Navarro for their helpful advise in early versions of manuscript and/or their technical support. A. Boyko and C. Bustamante (Cornell University) kindly provided resequencing information on *OR5I1*. Distributions of MAF and DAF proportions on 168 genes related to immune function and simulated data were kindly provided by P. Sabeti (Broad Institute). This research was supported by "Fundación Genoma España" (proyectos piloto CEGEN 2004–2005), the National Institute of Bioinformatics (http://www.inab.org), Dirección General de Investigación, Ministerio de Educación y Ciencia of Spain (grant BFU2005–00243), and Direcció General de Recerca, Generalitat de Catalunya (2005SGR00608). Andrés Moreno-Estrada is supported by a CONACYT fellowship from the Mexican government (grant 179339). SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (www.cegen.org).

Literature Cited

Aloni R, Olender T, Lancet D. 2006. Ancient genomic architecture for mammalian olfactory receptor clusters. Genome Biol. 7:R88.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics. 141:743–753.

Ben-Arie N, Lancet D, Taylor C, et al. (11 co-authors). 1994. Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. Hum Mol Genet. 3:229–235.

Bernstein HJ. 2000. Recent changes to RasMol, recombining the variants. Trends Biochem Sci. 25:453–455.

Cann HM, de Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. Science. 296:261–262.

Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 302:1960–1963.

Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J. 2004. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res. 32:W242–W248.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 131:479–491.

Finn RD, Mistry J, Schuster-Bockler B, et al. (13 co-authors). 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34:D247–D251.

Gardner M, Gonzalez-Neira A, Lao O, Calafell F, Bertranpetit J, Comas D. 2006. Extreme population differences across Neuregulin 1 gene, with implications for association studies. Mol Psychiatry. 11:66–75.

Gilad Y, Bustamante CD, Lancet D, Paabo S. 2003. Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet. 73:489–501.

Gilad Y, Man O, Paabo S, Lancet D. 2003. Human specific loss of olfactory receptor genes. Proc Natl Acad Sci USA. 100:3324–3327.

Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, Sharon D. 2000. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat Genet. 26:221–224.

Glusman G, Yanai I, Rubin I, Lancet D. 2001. The complete human olfactory subgenome. Genome Res. 11:685–702.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185:862–864.

Hall SE, Floriano WB, Vaidehi N, Goddard WA 3rd. 2004. Predicted 3-D structures for mouse I7 and rat I7 olfactory receptors and comparison of predicted odor recognition profiles with experiment. Chem Senses. 29:595–616.

Issel-Tarver L, Rine J. 1997. The evolution of mammalian olfactory receptor genes. Genetics. 145:185–195.

Katada S, Hirokawa T, Oka Y, Suwa M, Touhara K. 2005. Structural basis for a broad but selective ligand spectrum of a mouse olfactory receptor: mapping the odorant-binding site. J Neurosci. 25:1806–1815.

Malnic B, Godfrey PA, Buck LB. 2004. The human olfactory receptor gene family. Proc Natl Acad Sci USA. 101: 2584–2589.

Man O, Gilad Y, Lancet D. 2004. Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. Protein Sci. 13:240–254.

Mateu E, Calafell F, Ramos MD, Casals T, Bertranpetit J. 2002. Can a place of origin of the main cystic fibrosis mutations be identified? Am J Hum Genet. 70:257–264.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science. 310:321–324.

Nielsen R, Bustamante C, Clark AG, et al. (12 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30(38): 94–3900.

Reed RR. 1990. How does the nose know? Cell. 60:1–2.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70:841–847.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science. 312:1614–1620.

Safran M, Chalifa-Caspi V, Shmueli O, et al. (16 co-authors). 2003. Human Gene-Centric Databases at the Weizmann Institute of Science: geneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res. 31:142–146.

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234: 779–815.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576–1583.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin: A software for population genetics data analysis. Version 2.000. Geneva (Switzerland): Genetics and Biometry Lab, Department of Anthropology, University of Geneva.

Soranzo N, Bufe B, Sabeti PC, Wilson JF, Weale ME, Marguerie R, Meyerhof W, Goldstein DB. 2005. Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. Curr Biol. 15:1257–1265.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978–989.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Trask BJ, Massa H, Brand-Arpon V, et al. (11 co-authors). 1998. Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. Hum Mol Genet. 7:2007–2020.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Walsh EC, Sabeti P, Hutcheson HB, et al. (16 co-authors). 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. Hum Genet. 119:92–102.

Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D. 2004. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet. 74:637–646.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Anne Stone, Associate Editor

Accepted October 26, 2007