Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes

Yoav Gilad[1], Daniel Segré[1], Karl Skorecki[2], Michael W. Nachman[3], Doron Lancet[1] & Dror Sharon[1,4]

Substantial efforts are focused on identifying single-nucleotide polymorphisms (SNPs) throughout the human genome, particularly in coding regions (cSNPs), for both linkage disequilibrium and association studies[1,2]. Less attention, however, has been directed to the clarification of evolutionary processes that are responsible for the variability in nucleotide diversity among different regions of the genome[3]. We report here the population sequence diversity of genomic segments within a 450-kb cluster[4,5] of olfactory receptor (OR) genes[6,7] on human chromosome 17. We found a dichotomy in the pattern of nucleotide diversity between OR pseudogenes and introns on the one hand and the closely interspersed intact genes on the other. We suggest that weak positive selection is responsible for the observed patterns of genetic variation. This is inferred from a lower ratio of polymorphism to divergence in genes compared with pseudogenes or introns, high non-synonymous substitution rates in OR genes, and a small but significant overall reduction in variability in the entire OR gene cluster compared with other genomic regions. The dichotomy among functionally different segments within a short genomic distance requires high recombination rates within this OR cluster. Our work demonstrates the impact of weak positive selection on human nucleotide diversity, and has implications for the evolution of the olfactory repertoire.

A central aspect of research in human genetics concerns the association of variations in DNA sequences with human evolutionary history or heritable phenotypes. SNPs represent the most common type of variation, occurring on average once every 500–1,000 bp (ref. 8); however, there is considerable variability in the frequency of SNPs in different regions of the genome[9–11].

The OR proteins are G-protein-coupled receptors constituting the molecular basis for the sense of smell[7,12,13]. The OR genes, whose coding regions lack introns, are organized in genomic clusters[6,14]. We have analysed the population variability of seven intact OR genes, five pseudogenes and seven upstream intronic regions, all residing within a well-studied genomic cluster[4–6] on human chromosome 17p13.3 (Fig. 1).

We identified 59 SNPs (Table 1 and Fig. 2). By several measures, pseudogenes and introns had substantially more variation than intact genes (Table 1). Thus, the average number of haplotypes was significantly greater in pseudogenes (3.8±1.4) and introns (4.1±0.7; data not shown) than in intact genes (2.3±0.4; $Z=2.3$, $P<0.01$). Although the OR genes displayed a simple haplotype pattern, the OR pseudogenes and intron segments showed an unusual distribution (Fig. 2). Some, most notably *OR3A4P*, contained ancient allelic haplotypes in relatively high proportions, in conjunction with the most recent forms, with few or no intermediates (manuscript in preparation).

Both Watterson's θ and nucleotide diversity (π), were higher for the pseudogenes and introns than for the intact OR genes (Table 1). The average nucleotide diversity for pseudogenes was 0.11%, whereas average nucleotide diversity for introns was 0.10%, closely matching the average value for silent substitutions in the human genome[15]. In contrast, average nucleotide diversity for intact OR genes was only 0.03% (Table 1).

Several statistical tests are available for comparing the distribution of nucleotide variation in a population with that expected under a standard neutral model at equilibrium with respect to mutation and drift[16,17]. For example, Tajima's D is formulated so that the expected value of the test statistic is equal to 0 under the null hypothesis. Negative values for this statistic reflect an excess of low-frequency variants in the population, consistent with positive directional selection or a population expansion. Positive values reflect an excess of intermediate-frequency variants in the population, consistent with balancing selection or a population contraction. We observed nine loci with positive values and ten with negative values for Tajima's D, with average values near the neutral expectation of zero (Table 1). Fu and Li's D is similarly formulated, but is based on the number of singletons in a sample. Neither Tajima's D nor Fu and Li's D is significant for any of the loci (Table 1). Thus, there is no evidence for departures from neutral expectations in the frequency distribution, although it is important to bear in mind that these tests are not very powerful when few segregating sites are considered[18].

The average pairwise divergence between human and chimpanzee alleles for all segments is shown (Table 1). This analysis is based on comparisons between orthologues[19]. Divergence is greater at pseudogenes (D=2.34%) compared with genes (D=1.88%), although the difference in divergence is not as large

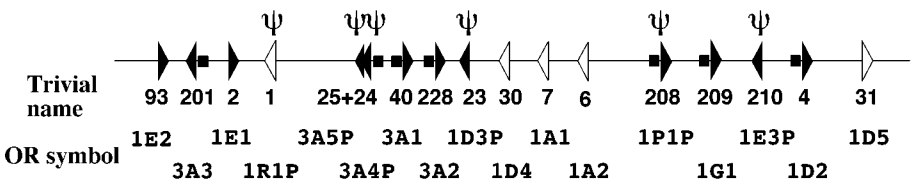

Fig. 1 The olfactory receptor gene cluster on human chromosome 17p13.3. Genes are depicted in the correct order, orientation and spacing within the cluster. The OR coding regions (black) are those resequenced for the presently study. ψ, Pseudogenes; squares, OR upstream introns from which an ~1-kb segment was resequenced. Two genes (*OR1A2* and *OR1A1*) and one pseudogene (*OR1R1P*) were not analysed due to problems in their PCR amplification from genomic DNA. Two more genes (*OR1D4* and *OR1D5*) were not included because they were 99.3% identical and their respective alleles could not be resolved.

[1]*Department of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot, Israel.* [2]*Molecular Medicine Laboratory and Department of Nephrology, Faculty of Medicine, Technion-Israel Institute of Technology, Bat-Galim, Haifa, Israel.* [3]*Department of Ecology and Evolutionary Biology, BioSciences West Bldg., University of Arizona, Tucson, Arizona, USA.* [4]*Present address: Ocular Molecular Genetics Lab, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to D.L. (e-mail: doron.lancet@weizmann.ac.il).*

Table 1 • Population variability parameters for intact OR genes, pseudogenes and introns

Gene	length	sample size	S	Watterson's θ	Nucleotide diversity	Tajima's D	Fu and Li's D	Average pairwaise D
OR1E2	972	34	1	0.03	0.04	1.52	–0.71	1.59
OR3A3	948	30	2	0.05	0.06	0.38	0.71	2.90
OR1E1	945	34	1	0.03	0.05	1.52	0.51	1.32
OR3A1	947	31	1	0.03	0.03	0.29	0.51	1.88
OR3A2	948	30	2	0.05	0.02	–1.27	–0.72	2.52
OR1G1	942	30	1	0.03	0.01	–1.00	–0.71	2.12
OR1D2	939	32	1	0.03	0.01	–1.24	0.51	0.85
average (genes)	948.7	31.57	1.3	0.04	0.03	0.03	0.57	1.88
s.d.	10.80	1.81	0.5	0.01	0.02	1.22	0.10	0.70
Pseudogene								
OR3A5P	771	32	2	0.07	0.06	–0.36	–0.72	2.26
OR3A4P	883	35	7	0.20	0.15	–0.73	1.23	1.81
OR1D3P	940	31	4	0.11	0.11	0.17	0.98	2.37
OR1P1P	993	31	3	0.08	0.10	0.75	0.86	2.05
OR1E3P	948	35	5	0.13	0.12	–0.35	0.98	3.19
average (pseudo)	907	32.8	4.2	0.12	0.11	–0.10	0.95	2.34
s.d.	85.50	2.05	1.9	0.05	0.03	0.57	0.19	0.52
OR introns								
OR3A3	1021	31	4	0.10	0.10	0.00	0.23	0.98
OR3A4P	968	30	8	0.22	0.22	0.12	0.78	1.12
OR3A1	914	32	5	0.14	0.13	–0.21	0.98	0.97
OR3A2	921	32	4	0.10	0.07	–0.89	0.23	3.75
OR1P1P	875	30	2	0.06	0.07	0.42	0.71	1.54
OR1G1	1134	31	3	0.07	0.06	–0.22	0.86	1.36
OR1D2	951	31	3	0.07	0.06	–0.56	0.86	1.32
average (introns)	969.1	31.00	4.1	0.11	0.10	–0.19	0.66	1.58
s.d.	86.04	0.82	2	0.06	0.06	0.43	0.30	0.98

More details on these genes are available (http://bioinfo.weizmann.ac.il/HORDE). Watterson's θ and the nucleotide diversity values are calculated per base pair in percentage.

as the difference in levels of human nucleotide diversity between these classes of loci. Notably, the observed divergence at OR pseudogenes (D=2.34%) is considerably higher than previous estimates for human-chimpanzee divergence at silent sites (D=1.4%; ref. 20) or pseudogenes (D=1.6%; ref. 21). OR pseudogene divergence, however, was nearly identical to silent-site divergence at intact OR genes (D=2.3%). This suggests that the underlying mutation rate may be similar in both classes of loci.

In general, population-level processes are expected to affect all loci in a roughly equal fashion, whereas deterministic processes, such as selection, are expected to act in a locus-specific manner. This general principle has led to several statistical tests for detecting selection, based on comparisons of levels of polymorphism with divergence for different loci[22,23]. The interspersed nature of OR genes and pseudogenes within a single genomic cluster on chromosome 17 presents an opportunity to disentangle selective effects acting on genes from demographic influences on all loci. Several observations in the current data set suggest that weak positive selection may be acting on intact OR genes.

First, the ratio of polymorphic to fixed variants for OR genes (9/114) versus OR pseudogenes (21/91) or OR introns (30/125) is significantly heterogeneous (Fisher's exact test (FET), $P \leq 0.01$ for each; Table 2). This deviation appears to be due largely to an excess of non-synonymous fixations. When non-synonymous sites within OR genes are compared with pseudogenes, the ratios are significantly heterogeneous (5/70 versus 21/91; FET, $P=0.01$), but when synonymous sites within OR genes are compared with pseudogenes, the ratios are not significantly different (4/44 versus 21/91; FET, $P>0.1$). The more conservative Hudson, Kreitman and Aguade (HKA) test[22] revealed a marginally significant reduction in the ratio of polymorphism to divergence at OR genes compared with pseudogenes ($0.05<P<0.10$).

Second, the ratio of non-synonymous to synonymous substitutions on a per-site basis for OR genes in a comparison between human and chimpanzee is high (average Ka/Ks=0.74; Table 3). For example, the average value of Ka/Ks for 49 genes in primates is 0.27 (ref. 24), approximately one-third the average value for OR genes. Ka/Ks values that are higher than average, but below unity, are difficult to interpret and may reflect either a low level of constraint or weak positive selection.

Third, HKA comparisons between either intact OR genes or OR pseudogenes and four other loci that have been well surveyed in humans (*DMD* (ref. 25), *PDHA1* (ref. 26), *LPL* (ref. 3) and *HBB* (ref. 27)) revealed significantly different ratios of polymorphism to divergence (Table 4). In all cases, loci in the OR cluster show lower levels of polymorphism relative to divergence than do the other genes. For each comparison, the effect is stronger for the OR genes than for the pseudogenes. These observations are consistent with weak positive selection depressing the level of variability in this region of chromosome 17, although differences in levels of polymorphism among loci may also be influenced by the different sampling schemes used in these studies.

It was an unexpected result that differences in patterns of variation occur among closely interspersed segments within a relatively small genomic region. The differences in levels of variation between genes and pseudogenes (Table 4) suggest that pseudogenes, which are typically 1–10 kb away from coding regions, undergo only modest 'hitchhiking' effects, whereby fixation of neutral mutations occurs through linkage to advantageous mutations. This, in turn, implies that there has been significant recombination in the history of our sample. In fact, the recombination rate for this region of chromosome 17, estimated from

Table 2 • Variation contrasts

	OR genes			OR pseudogenes	OR introns
	silent	replacement	total	total	total
SNPs	4	5	9	21	30
Fixed differences	44	70	114	91	125

Contrasts between human polymorphism and human-chimpanzee divergence in OR genes, pseudogenes and introns.

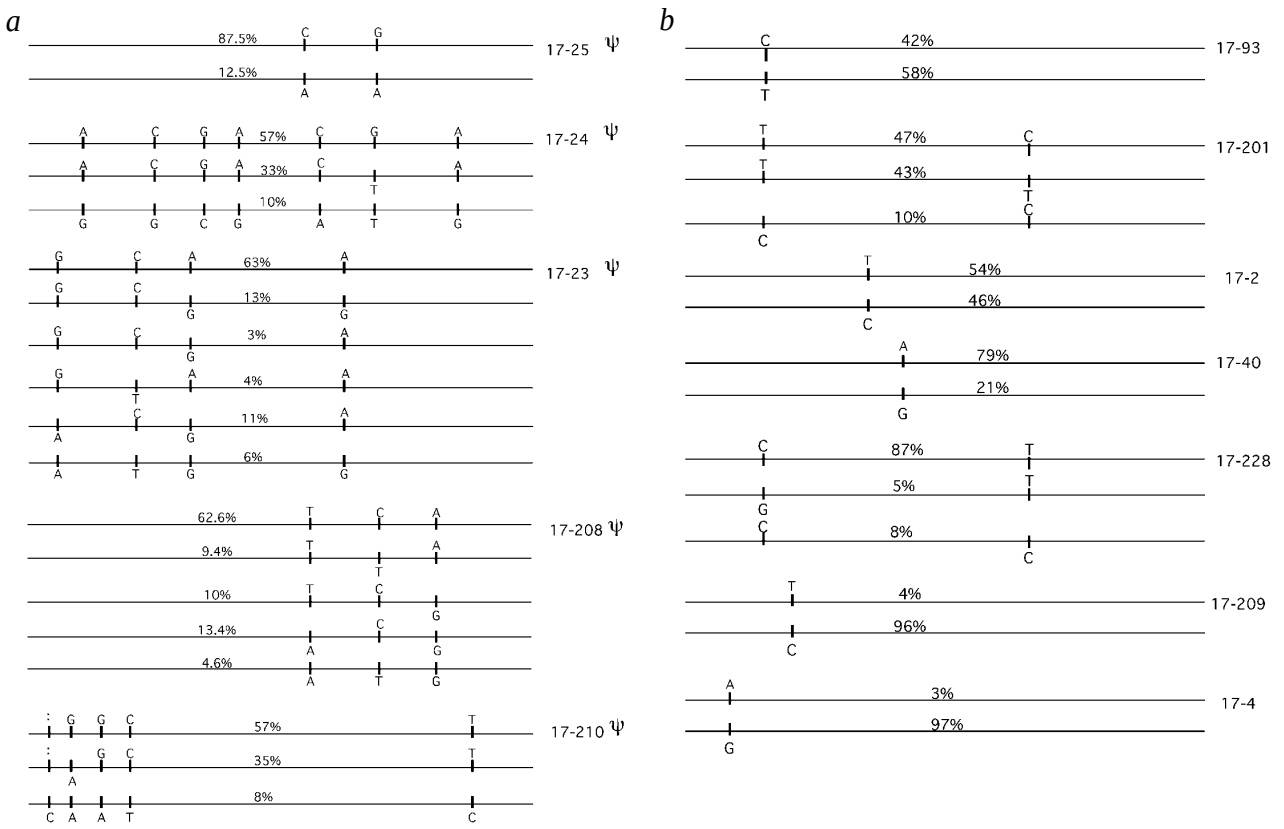

Fig. 2 Allelic haplotypes of OR genes and pseudogenes. Haplotypes of OR pseudogenes (***a***) and genes (***b***). The top nucleotides are the human-specific polymorphic bases and the bottom ones are the chimpanzee forms. ψ, A pseudogene. The per cent frequency of each allelic haplotype in the population is indicated.

comparison of genetic and radiation-hybrid maps[28], is 1.6 cM/Mb, somewhat higher than the genome-wide average rate of approximately 1 cM/Mb.

The signature of positive selection within the OR gene cluster is noteworthy. OR genes, likely constituting the largest multigene family in vertebrates, are under a unique set of evolutionary constraints and pressures. Individual OR genes may seldom be essential for survival, but it appears that the general enhancement and diversification of the size of the OR repertoire may confer a selective advantage. In addition, it has been suggested that for every odorant there may be several receptors that bind with different specificities[13]. Thus, a 'shifting terrain' scenario may be envisaged, in which relaxation of constraint occurs at some amino acid positions, whereas novel advantageous mutations are fixed under positive selection at other sites.

Our results suggest that weak positive selection can have important effects on the patterns of nucleotide diversity in human evolution. Studies of other gene families will help determine the generality of such findings.

Table 3 • Synonymous versus non-synonymous substitutions in intact OR genes

Gene	Ka (s.d.)	Ks (s.d.)	Ka/Ks
OR1E2	0.010 (0.006)	0.012 (0.007)	0.821
OR3A3	0.030 (0.009)	0.042 (0.020)	0.721
OR1E1	0.011 (0.007)	0.016 (0.012)	0.672
OR3A1	0.011 (0.006)	0.028 (0.011)	0.402
OR3A2	0.019 (0.008)	0.036 (0.018)	0.528
OR1G1	0.023 (0.009)	0.020 (0.012)	1.221
OR1D2	0.008 (0.004)	0.009 (0.006)	0.819
Average (sd)	0.016 (0.008)	0.023 (0.012)	0.740 (0.260)

Ka is the fraction of altering (non-synonymous) substitutions, Ks is the fraction of silent (synonymous) substitutions.

Methods

DNA samples. Human genomic DNA was from two sources. Genomic DNA of 20 unrelated anonymous individuals was isolated from buffy coats obtained from the Israeli Blood Bank, and samples from 30 unrelated individuals, provided by the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, were derived from the following ethnic groups: Ashkenazi Jews (n=10), Bedouins (n=10) and Yemenite Jews (n=10).

We isolated genomic DNA from chimpanzee (*Pan troglodytes*) from whole blood (provided by Y. Horvitz) using the Genomix DNA preparation kit (Talent SRL). Genomic DNA of chimpanzee was provided by K.K. Kidd.

PCR procedures. We designed primers (Table A, see http://genetics.nature.com/supplementary_info/) for PCR to amplify the full ORF of the 12 human OR genes, and seven ~1 kb segments from the OR introns, based on the available sequences[5]. We used the same primers for the sequencing reactions.

We carried out PCR in a total volume of 25 µl containing deoxynucleotides (0.2 µM of each; Promega), primers (50 pMol of each), PCR buffer (containing 1.5 µM $MgCl_2$, 50 µM KCl, 10 µM Tris, pH 8.3, 1 U Taq DNA polymerase (Boehringer)) and genomic DNA (50 ng). 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 and purified using the High Pure PCR Product Purification kit (Boehringer).

Cloning of PCR products. We subcloned the PCR products into the pAMP1 vector, without prior purification, using the Clone AmpTM System (Gibco BRL). We extracted the DNA from the subclones using the

Table 4 • HKA tests

Comparison	θ_1 (%)	θ_2 (%)	S_1/ES_1	S_2/ES_2	D_1/ED_1	D_2/ED_2	T	χ^2	P
DMD-OR	0.060	0.071	19/5.8	9/22.2	27/40.2	121/107.8	21.6	28.5	<0.001
PDHA1-OR	0.074	0.079	25/9.5	9/24.5	46/61.5	121/105.5	19.1	21.9	<0.001
LPL-OR	0.125	0.095	88/67.4	9/29.6	140/160.6	121/100.4	14.8	10.4	0.001
HBB-OR	0.079	0.076	28/13.5	9/23.5	30/44.5	121/106.5	20.2	19.2	<0.001
DMD-OR Ψ	0.092	0.147	19/8.9	21/31.1	27/37.1	101/90.0	12.63	10.6	0.001
PDHA1-OR Ψ	0.108	0.152	25/13.8	21/32.2	46/57.2	101/89.8	12.0	7.8	<0.01
LPL-OR Ψ	0.143	0.152	88/76.9	21/32.1	140/151.1	101/89.9	12.1	2.8	<0.10
HBB-OR Ψ	0.107	0.145	28/18.3	21/30.7	30/39.7	101/91.3	12.9	6.5	0.01

HKA tests comparing OR genes and OR pseudogenes with *DMD*, *PDHA1*, *LPL* and *HBB*. Subscripts denote locus 1 and locus 2, respectively. S, observed segregating sites; ES, expected segregating sites; D, observed divergence; ED, expected divergence; T, divergence time as in ref. 22.

Wizard Plus SV minipreps DNA purification system (Promega) and sequenced using vector primers from both directions.

DNA sequencing. Sequencing reactions were performed on PCR products or clones in both directions with dye-terminators (Dye terminator cycle sequencing kit; Perkin Elmer) on an ABI 373 or ABI 377 automated sequencer. After base calling with the ABI Analysis Software (version 3.0), the analysed data were edited using the Sequencher program (GeneCodes, version 3.0).

Determination of polymorphism and divergence. We determined haplotypes for each of the OR genes, introns and pseudogenes by sequencing PCR amplification products directly. For individuals containing more than one heterozygous site within a locus, multiple clones were sequenced to resolve phase. We used the Sequencher software to assemble the sequences and to identify DNA polymorphisms. To avoid sequencing errors, only SNPs appearing in two or more individuals were considered (7 singletons were excluded from data analysis which were found in *OR1E2*, *OR3A5P*, *OR3A2*, *OR1G1* and in the intronic segments of *OR3A3*, *OR3A4P*, *OR3A2*). Inclusion of these singletons in the analysis of polymorphism and divergence (Table 2) did not alter the results. The ratio of polymorphic to fixed variants for OR genes (12/114) versus OR pseudogenes (22/91) or OR introns (33/125) remains significantly heterogeneous (FET, $P<0.05$ for each). The human OR coding sequences were aligned with the chimpanzee and gorilla sequences.

Data analysis. The analyses described here depends on our ability to uniquely amplify products from genomic DNA that correspond to a single genetic locus. OR genes have been shown to undergo complex evolutionary processes of gene duplication and gene conversion[4,5,14,19]. Although more ancient gene duplications pose no problem, very recent ones may lead to ambiguity of allele identification, as exemplified by the gene pair *OR1D4* and *OR1D5* (ref. 5), which were excluded from our study. It is important to make sure that paralogues are not mistaken for alleles, as was done here, by ascertaining that in all individuals only a maximum of two alleles appear for each gene. Gene conversion would pose a barrier only if it is represented in polymorphic forms. Such instances have so far not been reported. After sequence alignment, the frequency of all polymorphic sites was calculated. We used two measures for the nucleotide variability in each locus; Watterson's θ (ref. 29), which is based on the number of segregating sites in the sample; and nucleotide diversity[30] (π), which is based on the average number of differences between all sequences in a sample. Tjima's D statistic[16] and Fu and Li's D (ref. 17) were calculated to test for deviations from neutral frequency distribution. Ratios of polymorphism to divergence were compared and assessed using Fisher's exact statistic test and the HKA test[22].

GenBank accession numbers. *OR1E1*, AF087916; *OR1D2*, AF087917; *OR1D3P*, AF087919; *OR3A4P*, AF087920; *OR3A5P*, AF087921; *OR3A1*, AF087924; *OR1E2*, AF087925; *OR3A3*, AF087926; *OR1P1P*, AF087927; *OR1G1*, AF087928; *OR1E3P*, AF087929; *OR3A2*, AF087930.

Acknowledgements

We thank U. Eisenstein for programming and B.-S. Boneh-Tamir for DNA samples. This research was supported by grants to D.L. from US National Institute of Health (DC00305), an Infrastructure grant of the Israel Ministry of Science, the German-Israel Foundation for Research and Development and the Krupp Foundation. And by the Israel Academy of Sciences-FIRST award to K.L.S. D.L. holds the Ralph and Lois Silver Chair in Human Genomics.

Received 8 May; accepted 15 August 2000.

1. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
2. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
3. Clark, A.G. *et al.* Haplotype structure and population genetic inference from nucleotide sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
4. Glusman, G., Clifton, S., Roe, R. & Lancet, D. Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity. *Genomics* **37**, 147–160 (1996).
5. Glusman, G. *et al.* Sequence, structure and evolution of complete human olfactory receptor gene cluster. *Genomics* **63**, 227–245 (2000).
6. Ben-Arie, N. *et al.* Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. *Hum. Mol. Genet.* **3**, 229–235 (1994).
7. Mombarts, P. Seven-transmembrane proteins as odorant and chemosensory receptors. *Science* **286**, 707–711 (1999).
8. Sherry, S.T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).
9. Przeworski, M., Hudson, R.R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302 (2000).
10. Nachman, M.W., Bauer, V.L., Crowell, S.L. & Aquadro, C.F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
11. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
12. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).
13. Lancet, D. Vertebrate olfactory reception. *Annu. Rev. Neurosci.* **9**, 329–355 (1986).
14. Trask, B.J. *et al.* Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **7**, 2007–2020 (1998).
15. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
16. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
17. Fu, Y.X. & Li, W.H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
18. Simonsen, K.L., Churchill, G.A. & Aquadro, C.F. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429 (1995).
19. Sharon, D. *et al.* Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. *Genomics* **61**, 24–36 (1999).
20. Hammer, M.F. A recent common ancestry for human Y chromosomes. *Nature* **378**, 376–378 (1995).
21. Miyamoto, M.M., Slightom, J.L. & Goodman, M. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* **238**, 369–373 (1987).
22. Hudson, R.R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).
23. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–654 (1991).
24. Ohta, T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**, 56–63 (1995).
25. Nachman, M.W. & Crowell, S.L. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy locus, Dmd, in humans. *Genetics* (in press).
26. Harris, E.E. & Hey, J. X chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
27. Harding, R.M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
28. Payseur, B.A. & Nachman, M.W. Microsatellite variation and recombination rate in the human genome. *Genetics* (in press).
29. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
30. Nei, M. & Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273 (1979).