# Selective pressures on the olfactory receptor repertoire since the human–chimpanzee divergence

**Alexander A. Gimelbrant*, Helen Skaletsky*†, and Andrew Chess*‡§**

*Whitehead Institute, †Howard Hughes Medical Institute, and ‡Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142

Edited by Masatoshi Nei, Pennsylvania State University, University Park, PA, and approved April 29, 2004 (received for review March 4, 2004)

**The availability of the sequence of the chimpanzee genome provides an opportunity to examine human genes and their chimpanzee orthologs and to analyze selective pressures that have been shaping the olfactory receptor repertoire since the human–chimpanzee divergence. We determined the ratio of nonsynonymous to synonymous changes for each of 186 orthologous pairs and then examined how the distribution of these ratios compares with the distribution expected under neutral drift. Consistent with the diminishing importance of olfaction for these species, we find no evidence for positive selection and we find evidence of weak purifying selection affecting over half of the repertoire.**

Olfaction has been important in the course of mammalian evolution, as witnessed by olfactory receptors' constituting the largest gene family (1) (1,000 genes or more). Knowledge of the selective pressures affecting the odorant receptor repertoire in various species is vital in understanding the corresponding importance of the olfactory system and has been of great interest (2–10). In mice, 80% of the identifiable olfactory receptor genes are intact (3, 4), whereas in primates, the repertoire is smaller, and as much as half the identifiable genes are clear pseudogenes (5–7); this observation suggests a diminished importance of olfaction in primates. On the other hand, given the different environments in which chimpanzees and humans have lived, an attractive hypothesis is that there has been positive selection on odorant receptors. Indeed, hints of such positive selection have been reported from studies that analyzed a relatively small number of genes (8–10). With the availability of the chimpanzee genomic sequence (www.nhgri.nih.gov/11509418) it is possible to assess the selective pressures through analyses performed on a large fraction of the odorant receptor repertoire.

## Methods

**Chimpanzee Olfactory Receptor Sequences.** The contigs with high-throughput shotgun genomic sequence of chimpanzee have been made available in GenBank by sequencing centers at Washington University and Massachusetts Institute of Technology (National Institutes of Health news advisory, www.nhgri.nih.gov/11509418). The chimpanzee olfactory receptor sequences were obtained by using BLASTN (11) with default parameters against the contig library by using full-length human olfactory receptor sequences (7) as queries, and then by ensuring that the pairs are reciprocal best hits. The set of olfactory receptor pairs with full-length ORFs was further refined by selecting only those chimpanzee sequences that would generate unique full-length alignments to the human genome assembly of July 2003, using at least 3 kb of chimpanzee genomic sequence including the olfactory receptor coding region. These alignments were generated by using the BLAT program (12) at the University of California Santa Cruz genome browser. A pair was rejected as possibly nonorthologous if the second-best alignment was 40% or more of the length of the best alignment. The complete set of the 186 orthologous olfactory receptor gene pairs used, with corresponding divergence statistics, is available as *Data Set 1*, which is published as supporting information on the PNAS web site.

The divergence statistics (Ka and Ks) were calculated by using the DIVERGE program within GCG [Wisconsin Package, v.10.3 (available from Accelrys, San Diego), using the method of Li *et al.* (13), as modified in refs. 14 and 15].

**Computer Simulations.** Perl scripts used for simulations are available on request. Briefly, the neutral drift model started with an olfactory receptor sequence (the results did not vary significantly with individual sequences used), and was subjected twice to random substitution at the rate of 0.65% [the empirically derived substitution rates for individual nucleotides were from table 1.5 in Li (16)]. For the two resulting sequences, diverged by 1.3% on average, Ka and Ka/Ks were calculated by using the GCG DIVERGE program. We conservatively assumed a divergence rate of 1.3%, slightly above the reported 1.23–1.24% rate (17, 18), as a higher divergence rate under neutral drift would result in fewer sequences with Ka/Ks differing from 1 in either direction. To simulate uniform purifying selection, a certain proportion of nonsilent substitutions resulted in the elimination of the sequence, as would occur during evolution: it would not be counted toward completion of the simulation experiment, and its divergence statistics would not be calculated. Parameter *c* (cumulative selection) defined the selective pressure such that at $c = 0.5$, in half the cases nonsynonymous substitutions would lead to sequence elimination. When used in a mixed model, only a given proportion of the sequences are subject to such uniform selection, whereas the rest are under neutral drift as described above.

## Results and Discussion

To examine the selective pressures that have been affecting the olfactory receptor repertoire since the chimpanzee-human divergence, we focused on unambiguous orthologous gene pairs with full-length coding regions. The nucleotide substitution rate between chimpanzee and human is slightly higher than 1.2% (17, 18). The overall low divergence, together with the availability of sequences flanking the coding regions, allows one to distinguish between orthologs and paralogs and therefore establish unambiguous orthologous relationships for a significant fraction of the olfactory receptor repertoire. Using sequences of intact human olfactory receptor genes (7) and the chimpanzee genomic sequence, we first looked for the reciprocal best sequence match for each coding region (see *Methods*). We further established orthology by selecting only those pairs that are contained within a 3-kb block of genomic sequence that can be aligned unambiguously between the two genomes. From this procedure we arrived at a set of 186 genes (see *Data Set 1*). This number is about one-half of 388, the recent estimate of the number of intact human olfactory receptor genes (7). Some orthologous olfactory receptor genes are likely to be in the yet-unavailable part of the chimpanzee genome sequence. In addition, we excluded from further analysis genes with ambiguous human–chimpanzee or-

This paper was submitted directly (Track II) to the PNAS office.

§To whom correspondence should be addressed. E-mail: chess@wi.mit.edu.

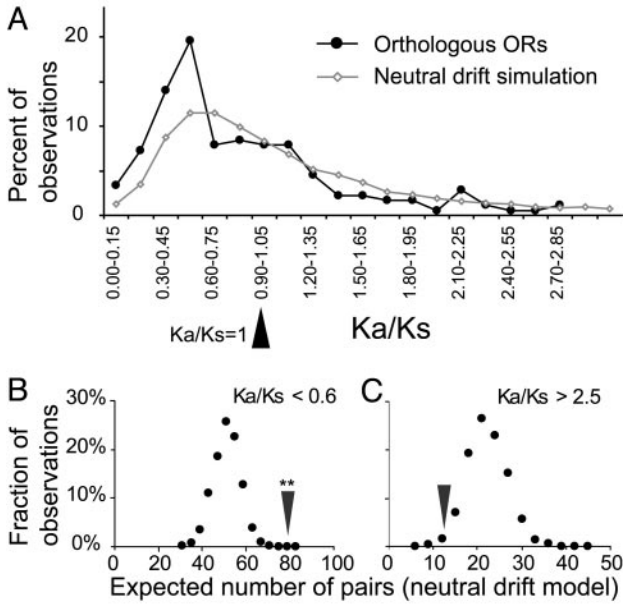


**Fig. 1.** Empirical distribution of Ka/Ks for olfactory receptor genes, compared with neutral drift simulation. (*A*) Distribution of Ka/Ks values for 186 full-length orthologous pairs of olfactory receptor genes (black curve) and distribution expected from neutral drift of olfactory receptor sequences (gray curve) assuming a uniform 1.3% divergence rate. (*B*) Estimation of the skewing of the empirical Ka/Ks distribution relative to the distribution expected under neutral drift. To estimate the probability that the deviation from the neutral drift distribution was caused by the limited size of the olfactory receptor set, expected number of sequences with Ka/Ks $< 0.6$ was determined from 1,000 sets of 186 sequences generated under the neutral drift model. The dots represent the resulting density histogram. The arrowhead denotes the empirical value for the set of olfactory receptors. The empirical value is expected with $P < 0.001$. (*C*) As in *B*, for Ka/Ks $> 2.5$.

thology, and apparent chimpanzee pseudogenes (some of which are likely to have resulted from sequencing errors). Genes that have undergone extensive gene conversion would also be excluded as lacking a sufficiently good match within coding sequence. The selected human sequences are distributed throughout the human genome.

Note that the presence of an uninterrupted ORF does not necessarily indicate that there is ongoing selection on a given gene, because even under completely neutral drift it takes time for a sufficient number of changes to accumulate to render a given gene an obvious pseudogene (for example, see ref. 19). Thus, for a gene family, if a fraction of the genes are intact this is not necessarily indicative of purifying selection on the family.

To investigate the selective pressures on the two repertoires, we took advantage of the large number of confirmed orthologous pairs and calculated Ka and Ks; Ka is the number of nonsynonymous (nonsilent) changes relative to the number of possible nonsynonymous changes for a particular sequence; and Ks is the number of synonymous (silent) changes relative to the number of possible synonymous changes for a particular sequence (13). For a given pair, a Ka/Ks $< 1$ indicates purifying selection and a Ka/Ks $> 1$ is consistent with positive selection. The black curve in Fig. 1*A* shows the distribution of Ka/Ks ratios for the 186 genes analyzed. The presence of a fraction of the repertoire with relatively high Ka/Ks ratios led us to examine the significance of the observed distribution.

We used a simulation to provide a basis for examining the distribution of Ka/Ks values for the repertoire. While any individual pairwise comparison is unlikely to be statistically significant, one can analyze the distribution for the repertoire as a whole and determine whether there is deviation from the distribution expected under the null hypothesis of neutral drift (lack of positive or negative selection). We used a simulation to generate the expected distribution under this null hypothesis. We simulated the effect of random nucleotide substitutions [using empirical nucleotide substitution rates (16)] on the Ks and Ka values of a representative olfactory receptor sequence as it would have diverged between human and chimpanzee (see *Methods*). We first confirmed that various olfactory sequences yielded similar results and then chose a single sequence and ran the simulation 100,000 times, yielding the gray curve in Fig. 1*A*.

When the observed data are compared with the simulation for the genes with Ka/Ks $> 1$, the overlap of the distributions suggests that the number of such genes is not greater than what would be expected by chance (Fig. 1*A*).

We also addressed the possibility that positive selection could be concentrated within certain portions of the coding region. To investigate this possibility we examined the frequency of amino acid changes in the 186 pairwise comparisons as a function of position in the coding region. We found that there is a relatively even distribution of amino acid changes (Fig. 4, which is published as supporting information on the PNAS web site) consistent with the idea that there is no positive selection on the repertoire since the chimpanzee–human divergence.

The comparison of the two distributions in Fig. 1*A* reveals that genes with a Ka/Ks ratio $< 0.6$ are more numerous than would be expected by chance (Fig. 1*A*). The Kolmogorov–Smirnov test (20) indicates that the two distributions are significantly different ($P < 0.001$). Thus, it appears that a fraction of the repertoire is under some purifying selection. To quantify the apparent shift of the observed distribution toward the lower Ka/Ks values, a simulation was performed in which groups of 186 pairs of sequences under completely neutral drift were sampled 1,000 times, and the expectation for number of sequences with Ka/Ks less than a threshold was compared with the empirically observed value. In the empirical set, 79 genes fall into the range of Ka/Ks $< 0.6$, when the expected number is $54 \pm 6.5$ (mean $\pm$ SD). This value or more would be expected only $<1/10{,}000$th of the time (Fig. 1*B*). A similar analysis of the orthologous pairs with Ka/Ks $> 2.5$ shows that the empirically observed value (12) is within the range of expected outcomes in the simulation (Fig. 1*C*).

As a further test of the selective pressures on the repertoire, we performed a comparative analysis of the observed and simulated data (under the neutral drift hypothesis), using only Ka instead of the Ka/Ks ratio (Fig. 2*A*). This analysis removes the variance contributed by the random fluctuation of the relatively low values of Ks. According to the Kolmogorov–Smirnov test, these distributions are significantly different ($P < 0.001$). Compared with the Ka distribution under neutral drift, there is an excess of sequence pairs with Ka between 0 and 0.008 (note that the genes in this Ka range have an average Ka/Ks ratio of 0.5, corresponding to apparent purifying selection).

To quantify the apparent shift of the empirical distribution toward the lower Ka values, the threshold analysis was performed as described above (Fig. 1 *B* and *C*) for Ka/Ks values. In the empirical set, 58 genes fall into the range of Ka $< 0.008$, when the expected number is $25 \pm 4.7$. This value or more would be expected only $<1/10{,}000$th of the time (Fig. 2*B*). A similar analysis of the orthologous pairs with Ka greater than 0.016 (with the average Ka/Ks ratio of 1.4, corresponding to apparent positive selection) shows that the empirically observed value (31) is within the range of expected outcomes in the simulation (Fig. 2*C*). Thus, analyses of Ka are consistent with purifying selection on a fraction of the repertoire. Moreover, there is no evidence for positive selection.

We next sought to estimate the fraction of the repertoire under purifying selection and, at the same time, to estimate the strength

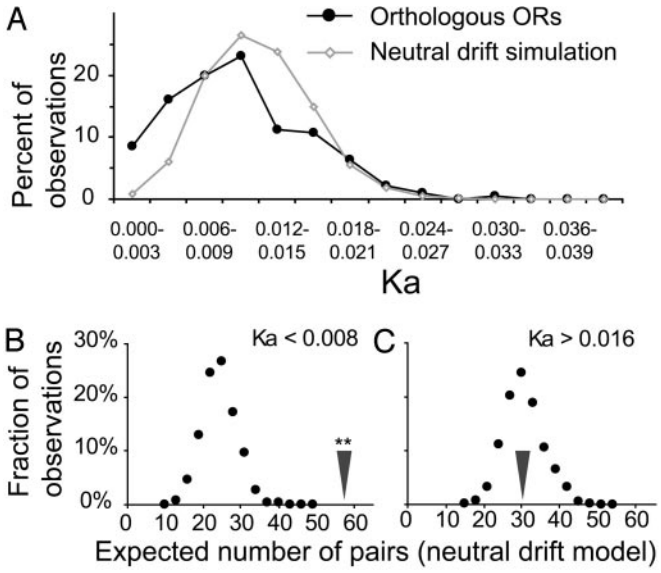


**Fig. 2.** Empirical and simulated distributions of Ka values of olfactory receptor genes. (*A*) Distribution of Ka values for 186 full-length orthologous pairs of olfactory receptor genes (black curve) and distribution expected under neutral drift of olfactory receptor sequences (gray curve) assuming a uniform 1.3% divergence rate. (*B*) The expected number of sequences with Ka $<$ 0.008 was determined from 1,000 sets of 186 sequences generated under the neutral drift model. The dots represent the resulting density histogram with bin size of 3. The arrowhead denotes the empirical value for the set of olfactory receptors; ** denotes $P < 0.001$. (*C*) As in *B*, for Ka $>$ 0.016.

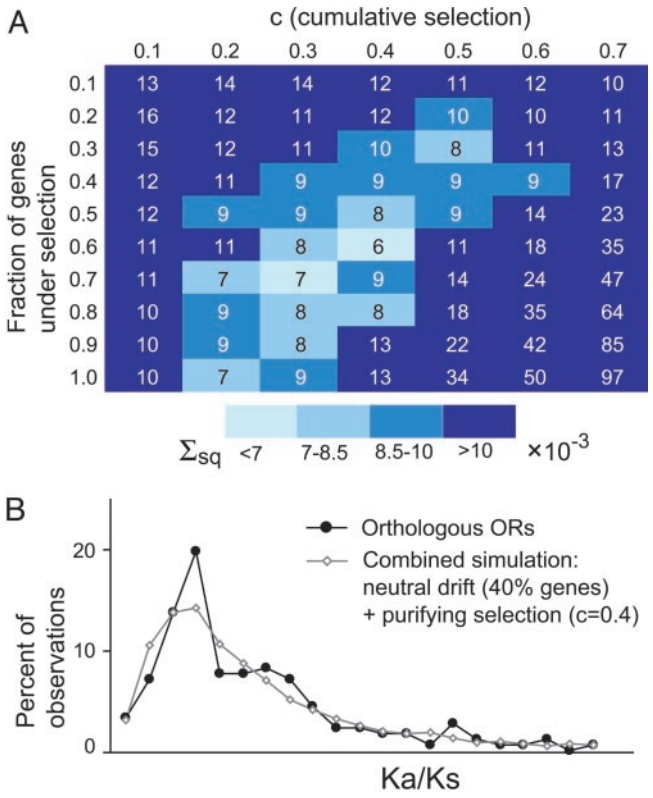


**Fig. 3.** Weak purifying selection approximates observed Ka/Ks distribution. (*A*) Fitting simulated Ka/Ks distributions to the empirical one. Shown is sum of squares of deviations ($\Sigma_{sq}$) from the observed Ka/Ks distribution for the simulated distributions, depending on the strength of cumulative purifying selection (*c*) and the fraction of the genes under selection. The rest of the genes are under neutral drift. (*B*) Comparison of the distribution of Ka/Ks values for 186 full-length orthologous pairs of olfactory receptor genes (black curve, same as in Fig. 1) with the distribution resulting from a simulation assuming 60% of genes under weak ($c = 0.4$) cumulative purifying selection and 40% under neutral drift (gray curve).

of the selection. We performed an analysis that involved generating simulations based on varying two parameters: first, the fraction of the repertoire under uniform purifying selection (with the rest of the repertoire under a model of neutral drift); and second, the strength of the purifying selection. We then determined which parameters yield simulation-generated data that most closely approximate the observed data.

Absent knowledge of the fraction of the repertoire that is under purifying selection, we tested 10 different fractions (ranging from 0.1 to 1.0), and the remaining sequences were allowed to diverge under no selection (as was done in the first simulation). For each assumed fraction of the repertoire under selection, we performed seven distinct simulations (5,000 trials per simulation) representing seven different (uniform) strengths of purifying selection. The resulting Ka/Ks ratio distributions were then compared with the observed distribution. Closely matched distributions therefore reveal plausible combinations of the strength of the selection and the fraction of the repertoire subject to selection (Fig. 3*A*). For these combinations, as the fraction of the genes under selection increases, the value of the selective pressure decreases. These analyses suggest that it is likely that between 50% and 90% of the orthologous pairs appear to be under weak purifying selection ($0.2 \leq c \leq 0.4$, where $c$ is cumulative selection, the probability that a given nonsynonymous change will be eliminated by selection in the time since the chimpanzee–human divergence). As an example, a curve that fits the observed data quite well is shown in Fig. 3*B*. There 60% of genes are under weak purifying selective pressure ($c = 0.4$). The Kolmogorov–Smirnov test indicates that these two distributions are not significantly different.

Analyzing unambiguously orthologous olfactory receptors from the chimpanzee and human genomes, we have provided evidence consistent with a predominance of purifying selection affecting over half the repertoire. Moreover, orthologous pairs that at first glance could be construed to have been under positive selection are no more prevalent than what would be expected by chance under models with no positive selection. While selective pressures on individual genes are likely varied rather than uniform, our analyses suggest that olfactory receptor genes under strong positive or purifying selection are at most a small minority in the repertoire.

A prior examination of the selective pressures affecting the olfactory receptor repertoire (since the chimpanzee–human divergence) claimed evidence of positive selection on the olfactory receptor repertoire since the chimpanzee–human divergence (10). However, the statistical approach used in that study has been called into question as prone to detection of false positives (21, 22). Apart from this possibility, the contrast between our findings and the prior report (10) is likely due to our analysis of a much larger set of genes and our more reliable ascertainment of the orthology. We analyzed 186 pairs of apparently intact olfactory receptors, as opposed to 46 olfactory receptors, 11 of which were apparent pseudogenes in humans. Furthermore, our use of the flanking sequence surrounding the olfactory receptor coding regions to firmly establish orthologous pairs eliminated the confounding effects of the paralogs. Other prior studies analyzed even smaller numbers of orthologous intact olfactory receptors (8, 9).

Our demonstration that there is no evidence of positive selection on the olfactory receptor repertoire as a whole since the chimpanzee–human divergence must be viewed in the context of other observations about the repertoire in these two

species and other primates. Primates with trichromatic color vision have a larger fraction of pseudogenes than other primates and other mammals, suggesting that olfaction is less important to these species (2). The presence of a large (>1,000 members) repertoire in mammals (1) indicates that the expansion and maintenance of this repertoire was, at some time in evolution, under considerable positive selection. It is perhaps not surprising that there is a lack of positive selection in the time since the human–chimpanzee divergence because the numbers of apparently intact olfactory receptor genes in these species appear to be shrinking. Our findings are consistent with a diminished importance of the olfactory system in humans and chimpanzees, and consequent loss or relaxation of selective constraints on the olfactory receptor genes.

We thank A. Bortvin, M. Daly, A. Ensminger, A. Y. Gimelbrant, D. Page, S. Rozen, S. Schaffner, and P. Sklar for comments. This work was supported by grants from the National Institutes of Health (to A.C.).

1. Buck, L. & Axel, R. (1991) *Cell* **65,** 175–187.
2. Gilad, Y., Wiebe, V., Przeworski, M., Lancet, D. & Pääbo, S. (January 20, 2004) *PLoS Biol.* **2,** 10.1371/journal.pbio.0020005.
3. Zhang, X. & Firestein, S. (2002) *Nat. Neurosci.* **5,** 124–133.
4. Young, J. M., Friedman, C., Williams, E. M., Ross, J. A., Tonnes-Priddy, L. & Trask, B. J. (2002) *Hum. Mol. Genet.* **11,** 535–546.
5. Glusman, G., Yanai, I., Rubin, I. & Lancet, D. (2001) *Genome Res.* **11,** 685–702.
6. Zozulya, S., Echeverri, F. & Nguyen, T. (June 1, 2001) *Genome Biol.* **2,** Research0018.1–0018.12.
7. Niimura, Y. & Nei, M. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 12235–12240.
8. Gilad, Y., Segre, D., Skorecki, K., Nachman, M. W., Lancet, D. & Sharon, D. (2000) *Nat. Genet.* **26,** 221–224.
9. Gilad, Y., Bustamante, C. D., Lancet, D. & Pääbo, S. (2003) *Am. J. Hum. Genet.* **73,** 489–501.
10. Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., *et al*. (2003) *Science* **302,** 1960–1963.
11. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
12. Kent, W. J. (2002) *Genome Res.* **12,** 656–664.
13. Li, W. H., Wu, C. I. & Luo, C. C. (1985) *Mol. Biol. Evol.* **2,** 150–174.
14. Li, W. H. (1993) *J. Mol. Evol.* **36,** 96–99.
15. Pamilo, P. & Bianchi, N. O. (1993) *Mol. Biol. Evol.* **10,** 271–281.
16. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
17. Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S. F., Park, H. S., Yaspo, M. L., Lehrach, H., Chen, Z., *et al*. (2002) *Science* **295,** 131–134.
18. Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. (2002) *Am. J. Hum. Genet.* **70,** 1490–1497.
19. Zhang, J. & Webb, D. M. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8337–8341.
20. Fisher, L. D. & Belle, G. V. (1993) *Biostatistics: A Methodology for the Health Sciences* (Wiley, New York).
21. Suzuki, Y. & Nei, M. (2004) *Mol. Biol. Evol.* **21,** 914–921.
22. Zhang, J. (March 19, 2004) *Mol. Biol. Evol.*, 10.1093/molbev/msh117.