Human Molecular Genetics, 2002, Vol. 11, No. 5 535–546

Different evolutionary processes shaped the mouse and human olfactory receptor gene families

Janet M. Young, Cynthia Friedman, Eleanor M. Williams, Joseph A. Ross, Lori Tonnes-Priddy and Barbara J. Trask*

Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., C3-168, Seattle, WA 98109, USA

Received November 2, 2001; Revised and Accepted December 20, 2001 DDBJ/EMBL/GenBank accession nos BH405737–BH406512

We report a comprehensive comparative analysis of human and mouse olfactory receptor (OR) genes. The OR family is the largest mammalian gene family known. We identify ~93% of an estimated 1500 mouse ORs, exceeding previous estimates and the number of human ORs by 50%. Only 20% are pseudogenes, giving a functional OR repertoire in mice that is three times larger than that of human. The proteins encoded by intact human ORs are less highly conserved than those of mouse, in patterns that suggest that even some apparently intact human OR genes may encode non-functional proteins. Mouse ORs are clustered in 46 genomic locations, compared to a much more dispersed pattern in human. We find orthologous clusters at syntenic human locations for most mouse genes, indicating that most OR gene clusters predate primate–rodent divergence. However, many recent local OR duplications in both genomes obscure one-to-one orthologous relationships, thereby complicating cross-species inferences about OR–ligand interactions. Local duplications are the major force shaping the gene family. Recent interchromosomal duplications of ORs have also occurred, but much more frequently in human than in mouse. In addition to clarifying the evolutionary forces shaping this gene family, our study provides the basis for functional studies of the transcriptional regulation and ligand-binding capabilities of the OR gene family.

INTRODUCTION

Mammals are able to detect and discriminate thousands of different odors (1). This capability is important to find food, identify mates and offspring, and avoid danger. The first step in the complex pathway resulting in the sense of smell is the interaction of odorant molecules with olfactory receptors (ORs) in the nose. ORs are G-protein-coupled seven-transmembrane-domain proteins that can trigger a signaling cascade in sensory neurons (2). Recognition of diverse odorants is achieved by using an estimated 1000 OR genes distributed around the rodent (3) and human genomes (4). It is the largest mammalian gene family known, comprising 1/30 to 1/50 of all genes in the genome. However, the evolution, transcriptional regulation and odorant binding capabilities of the OR gene family are still poorly understood.

Over 900 human OR genes were identified recently in the almost complete human genome sequence (4). Approximately 350 of these genes are intact and appear to be functional (5). Most human OR genes are clustered in the genome in arrays that can contain over 100 genes (4). Human OR genes have been found at over 40 locations in the human genome by fluorescence *in situ* hybridization (FISH) (6,7) and at over 100 locations by sequence analysis (4). The human regions containing OR genes show a bias for chromosomal bands near telomeres and centromeres (6,7). The human OR gene family appears to be evolving quickly—around half of human OR-containing genomic clones hybridize to more than one genomic location, indicating that large blocks of DNA containing these genes have duplicated recently (7). Some of these duplications are so recent that their copy number is polymorphic in the human population (8).

Individual humans vary in their ability to detect some odors, and some specific anosmias have been shown to be genetically determined (9,10). In no case has the molecular basis of variation or deficit in sensory perception been defined. Odorant–receptor relationships are not yet known for any human gene. Functional studies of human OR genes are hindered by the difficulties encountered in attempts to obtain live neurons from human donors (11) and to functionally express OR genes in heterologous cell lines (12). Studies in experimentally tractable model organisms, such as mouse, will be needed to determine the ligand-binding properties of OR genes and to understand how these genes are regulated. The identification of orthologous relationships between human and mouse OR genes will be key to translating data from mouse studies into an understanding of human olfaction. So far, the comparative analysis of only a few pairs of mouse and human orthologous clusters has been reported (13–17).

The murine OR gene family is much less well characterized than the human OR family. In one study, 21 OR genes were

*To whom correspondence should be addressed. Tel: +1 206 667 1470; Fax: +1 206 667 4023; Email: btrask@fhcrc.org

Present address:
Lori Tonnes-Priddy, Epigenomics Inc., 1000 Seneca Street, Suite 300, Seattle, WA 98101, USA

found at 11 different genomic locations (18), and various other studies have identified additional loci (19–21). Analysis of small samples of OR genes suggests that most mouse genes are functional, whereas a substantial fraction of OR genes in microsmatic species such as hominoids, old world monkeys and dolphins are pseudogenes (22,23). Genomic sequences at several mouse OR loci have been recently characterized (13,14,16,24), but these studies give only a limited picture of the gene family. Knowledge of the entire gene family will provide the basis for studies of transcriptional regulation and receptor–ligand interactions in the mouse.

The recent availability of a whole genome shotgun sequence of the mouse (Celera Genomics) has enabled us to assemble an almost complete catalog of OR genes. With this catalog, we describe the evolution of the OR genes and report striking differences between the human and mouse OR gene families with respect to pseudogene content, protein sequence conservation and mechanisms of duplication. The differences we observe between orthologous gene clusters give insights into the pressures and processes that have acted on this gene family during rodent and primate evolution.

RESULTS

Human OR genes are more dispersed in the genome than mouse ORs

In order to determine the genomic distribution of the mouse OR gene family, we identified and mapped OR-containing bacterial artificial chromosome (BAC) clones. BAC clones covering 1.6% of the mouse genome (2471 clones) were positive in a hybridization screen for OR genes using probes made by degenerate PCR. A subset of 94 BACs was subjected to secondary PCR tests; 94% of these clones were confirmed to contain OR genes. Thus, only a small proportion of clones may be false positives due to low-stringency hybridization conditions.

FISH of 272 of the hybridization- and PCR-positive clones (approximately 1.3× clone coverage of the OR subgenome) shows that there are at least 32 cytogenetically distinct OR-containing locations in the mouse genome (Fig. 1), although additional loci can be identified by sequence analysis (see below). Half of the 272 FISH-mapped clones were chosen randomly and the remainder were chosen to ensure good coverage of OR-containing genomic regions (Materials and Methods). End sequences were obtained for the FISH-mapped clones to allow integration of cytogenetic and sequence information and independently verify the map location of OR clusters found by sequence analysis (see below). The number of mouse OR locations found by FISH is fewer than the 42 locations previously observed by FISH in the human genome (7), despite the fact that far more mouse OR clones than human OR clones were analyzed. This result indicates that the human OR gene family is more dispersed than the mouse family. The distribution of locations is also grossly different; only 8/32 (25%) of mouse OR locations are in subtelomeric or pericentromeric bands, compared to 23/42 (55%) of human locations (7).

FISH analysis can also detect recent duplications of OR-containing blocks to cytogenetically distinct chromosomal locations that might not be apparent from sequence analysis. Our FISH results show that such interchromosomal duplication events are less frequent in the mouse genome than in the human genome. Only 4% (10/272) of the mouse OR-containing clones resulted in FISH signals at two or more genomic locations (Fig. 1), while around half of human OR-containing clones do so (7). Multiple hybridization signals indicate recent large duplication(s) involving at least some of the sequence contained in the clone.

The mouse genome contains approximately 1500 OR genes

By searching Celera's mouse genome assembly, we have identified 866 intact, full-length OR genes and 340 apparent pseudogenes. Partial sequence data are available for an additional 187 genes, making a total of 1393 OR sequences. Our database-searching strategy used 34 OR protein queries to search the mouse genome (Materials and Methods). Our original blast searches were sensitive enough to find—and eliminate from further analysis—95 sequences that matched a non-olfactory G protein-coupled receptor better than an OR. We are reasonably confident that our analysis is restricted to bona fide OR genes; the characteristic protein sequence motifs of this gene family are remarkably well conserved in the genes identified (see below). In addition, although intact sequences were not selected with a percent identity-based cutoff, all 866 intact genes share ≥40% amino acid identity with an annotated OR sequence from the public databases. To date, we have experimentally validated over 400 of the identified OR genes by sequencing cDNA clones derived from mouse olfactory neuroepithelium (J.Young, J.Ross, E.Williams, T.Newman, L.Tonnes-Priddy, R.Lane and B.Trask, manuscript in preparation).

Two subsets of the OR genes were analyzed further. The 'full-length' dataset comprises the 1054 sequences of both genes and pseudogenes, but not sequences interrupted by repeats or by ends or gaps in scaffold sequences. The 'comprehensive' dataset contains all 1468 OR sequences identified, including all partial sequences, some of which are redundant with one another (i.e. they represent short scaffold sequences, which should have assembled with other short scaffolds or into gaps in the larger scaffolds).

In order to assess the sensitivity of our method of database mining and to estimate the coverage and sequence error rate of the OR subgenome in the Celera assembly, a non-redundant set of all 155 previously identified mouse OR nucleotide sequences was downloaded from GenBank (25). Of these 155 sequences, 143 (93%) match a sequence in the comprehensive dataset with ≥98% identity over ≥200 bp. Therefore, we estimate that the complete mouse genome contains at least 1510 OR sequences (1393 ÷ 0.93).

Failure to find 12 Genbank OR sequences is not due to insensitivity in our OR gene-finding method. When these sequences were used to search the entire Celera mouse genome assembly, none of the 12 genes was present. One of the non-matching Genbank sequences (GenBank accession no. X89682) is mislabeled as a mouse sequence; it must be of human origin, since it exactly matches several human genomic sequences in Genbank.

The mouse OR gene family has 20% pseudogenes

Of the 340 apparent pseudogenes in the comprehensive dataset that are not interrupted by gaps in the sequence data, 134 (39%) are interrupted by interspersed repeat sequences, 27 are not interrupted by any recognizable repeat, but do not align to

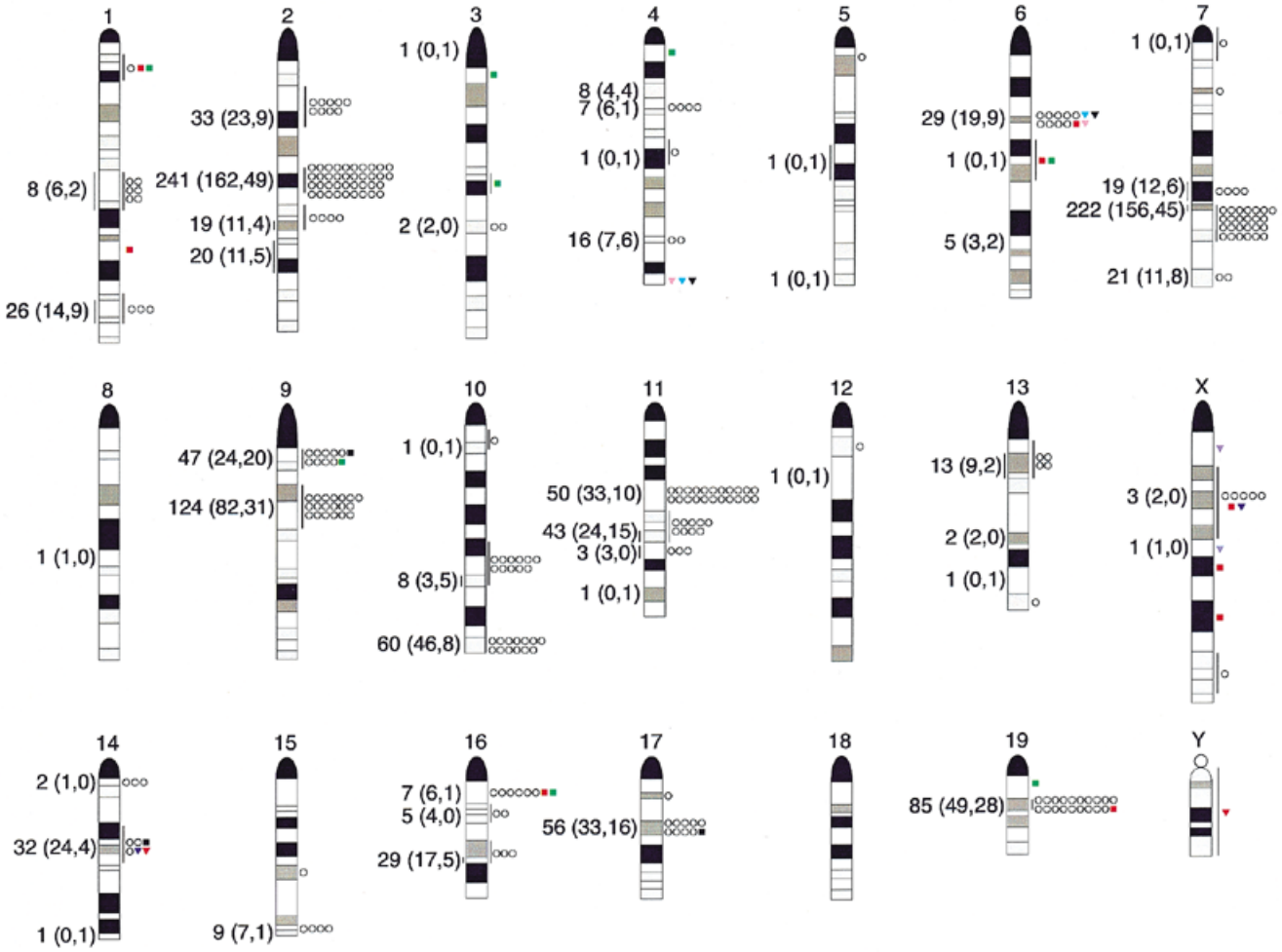

Figure 1. An ideogram showing the locations of 1267 mouse OR genes. Symbols to the right of the chromosomes indicate the locations of 272 mouse OR-containing BACs as determined by FISH. Each open circle represents a BAC that hybridized to only one location. Other symbols represent BACs hybridizing to two (triangles) or more (squares) genomic locations, and each color represents the same BAC. Vertical lines at some locations indicate that these bands were not distinguishable at the resolution used for FISH. Numbers on the left side of each chromosome indicate the number of OR gene sequences found at each location by data-mining the Celera whole genome shotgun assembly. The two numbers in parentheses indicate the number of genes known to be intact or pseudogenes, respectively (genes with incomplete data are not tallied). Scaffolds containing 1267 of the 1468 OR sequences identified in the Celera assembly were mapped using data from two complementary sources. In total, 79 scaffolds were mapped using one or both data sources. A chromosomal location was assigned to 65 of the scaffold sequences via matches to end sequences of the FISH-mapped BACs. Celera reports chromosomal locations for 75 OR-containing scaffolds based on linkage/radiation hybrid mapping data for markers in the sequences. Unmapped scaffolds are short, typically spanning <2 kb. In 54/60 cases where information is available from both Celera and FISH-based mapping, there is good agreement. In the six discordant cases, we used our own results when several FISH-mapped BACs were in agreement (three cases; two involving sequences mapped by Celera to 11B1 and by us to 2E1). In two discrepant cases, only one BAC was FISH-mapped; we consider these scaffolds unmapped. The remaining discordancy appears to result from a false join in scaffold GA_x5J8B7W3KVV; five BACs whose end sequences match near one end of the scaffold FISH-map to 2E1, and five BACs matching the other end FISH-map to 11B1-11B2. The prevalence of discordancies involving 11B1 and 2E1 (three of the six cases) suggests a systematic error in Celera's map for chromosome 11B1.

other ORs over their entire length, and the remaining 179 are full length, but contain one or more stop codons and/or frameshifting errors.

Based on Celera sequence, 28% of OR sequences appear to encode pseudogenes, but some of the full-length pseudogenes are likely to be intact genes with sequencing errors, since the Celera assembly is still in draft form. To estimate the rate of frameshifting sequencing errors, we compared Celera and Genbank sequences for the 123 Celera sequences (including 15 pseudogenes) that matched a Genbank OR with ≥99.5% identity. Six of these Celera sequences have one or more single base pair insertions or deletions in their coding region as compared to the Genbank sequence. In all six cases, the Celera sequence appears to be a pseudogene, and the Genbank sequence appears intact, strongly suggesting that the discrepancy is due to an error in the Celera sequence (we encountered nine frameshift errors in 93 kb sequence surveyed). Given this error rate, we estimate that approximately 70 of the apparent pseudogenes are actually intact, yielding approximately 940 intact genes and 250 pseudogene sequences, or a pseudogene fraction of 20%. If this 20% rate is applied to our whole-genome estimate, we extrapolate a total of approximately 1210 intact genes and 300 pseudogenes.

Mouse OR clusters map to 46 genomic locations

The 1468 OR genes in the comprehensive dataset derive from 243 of Celera's scaffold sequences, reflecting the clustered

organization of these genes in the genome (see below). Scaffolds containing 1267 (86%) of these genes could be assigned to 46 genomic locations using two complementary methods (Fig. 1). End sequences of FISH-mapped BAC sequences were used to map 65 scaffold sequences; Celera used the genetic or radiation-hybrid map positions of markers in the sequence to map 75 scaffold sequences (details in legend to Fig. 1). The number of mouse OR locations is far fewer than the 104 OR-gene locations in the human genome (4). We have identified OR sequences in Celera scaffolds corresponding to every OR location detected by FISH with at least two clones. Sequence analysis uncovered an additional 12 OR-containing sites not detected by our FISH analysis of 272 BACs (see above). Missing locations are expected, since the clones analyzed by FISH represent only a 1.3-fold coverage of the OR subgenome. Of these 12 sites, eight contain only one OR gene.

OR genes are arranged in the mouse genome in clusters containing an average of 16 genes with average gene-to-gene spacing of 21 kb. We examined the spacing between all genes identified (comprehensive dataset) and found that the distance between neighboring genes on the same scaffold varies considerably, from 318 bp to >5 Mb, although 90% of distances are <40 kb. In the eight cases where gene-to-gene distance is >0.5 Mb, it appears that two distinct clusters are present in the same scaffold sequence (the spacing between the two clusters is much more than the average spacing within the clusters). Using 0.5 Mb as the cutoff distance for distinguishing OR clusters, the average gene-to-gene spacing within clusters is 21 kb, but is highly variable (SD = 26 kb). Gene-to-gene distances may partly reflect the requirement for space upstream of genes for 5′ untranslated exons and transcriptional control regions. Only 10/40 (25%) of genes with another OR gene <5 kb upstream have full length sequence available and are apparently intact, compared to 59% for all genes, suggesting that genes without these upstream sequences degenerate into pseudogenes.

To estimate cluster size, we considered only OR sequences on Celera scaffolds spanning over 1 Mb, as these are more likely to contain complete OR clusters than are shorter scaffolds. There are 72 such clusters containing a total of 1164 genes in the comprehensive dataset. Of these 'clusters', 20 contain only one gene (1.7% of genes), but most genes (1018 or 87%) are in clusters of 10 or more genes. In contrast, 50 human genes are in singleton 'clusters' (4), reflecting the greater genomic dispersion of the human gene family. Our ability to determine cluster size is limited by gaps between Celera scaffold sequences; it will therefore be an underestimate of true cluster size. With this caveat, cluster size ranges from 1 to 98 genes (mean = 16) and is highly variable (SD = 22). The physical size of clusters is also very variable and ranges from 910 bp (one gene) to ~2 Mb (mean = 340 kb; SD = 470 kb).

OR proteins are less conserved in human than mouse

Alignment of protein translations of the 866 intact mouse OR genes reveals highly conserved motifs in some regions of transmembrane domains (TM) 2, 3, 6 and 7, as well as at several extracellular cysteine residues and some other small motifs (e.g. S-Y in TM5). Other positions in the protein are highly variable. Three positions are absolutely conserved in all 866 mouse sequences, and there are 16 positions where ≥99% of proteins have the same amino acid. These positions are less conserved in the 347 intact human OR genes reported by Zozulya *et al.* (5). There are no absolutely conserved positions in the human proteins, and only two positions where ≥99% of proteins have the same amino acid. Many other positions show lower conservation in the human proteins than in mouse. For example, an arginine residue is found in the conserved MAYDRYVAIC motif (TM3) in 98% of mouse sequences, but only 89% of human sequences (Fig. 2). Sequence conservation can also be measured using information theory, where the 'information content' of each position in the sequence is scored on the basis of the distribution of amino acids present (26), with conserved positions scoring more than variable positions. The total information content of the mouse and human proteins are 668.6 ± 0.2 bits and 645.6 ± 1.1 bits, respectively, confirming that the human ORs are less conserved than the mouse family.

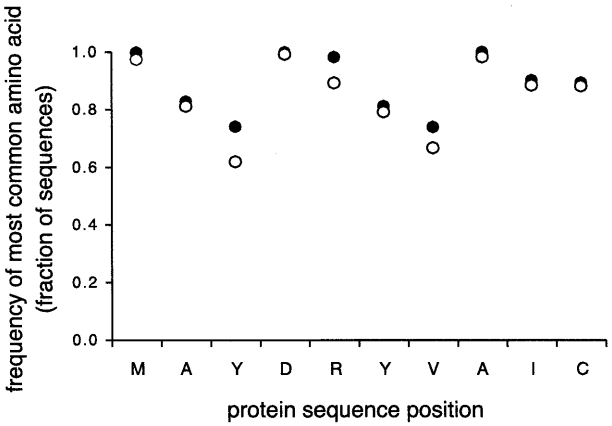

Figure 2. Intact mouse ORs are more conserved than human ORs. Frequency of the most common amino acid at each position in the conserved 'MAYDRYVAIC' region of TM3. Frequencies in the mouse gene family are plotted as closed circles and frequencies in the human gene family are plotted as open circles. A total of 866 mouse and 347 human genes were evaluated.

Recent tandem events have shaped the OR gene family

An alignment of all human and mouse OR genes shows that genes near each other in the genome are often very similar in sequence, implying that tandem events (duplications and/or gene conversions) are the major evolutionary force shaping the diversity of this gene family (Fig. 3). For 823 (78%) of the 1054 mouse genes in the full-length dataset, the closest mouse relative resides in the same genomic cluster. In the human genome, tandem events are also the major force, with 484 (73%) of 665 full-length genes related most closely to another gene in the same cluster. These tandem duplications are also evident on a phylogenetic tree (Fig. 4).

However, a subset of human OR genes has recently duplicated interchromosomally, resulting in highly similar genes in distant genomic locations (Fig. 3B, arrow). There are 203 pairs of human genes that share >90% amino acid identity. Of these 203 very similar gene pairs, 120 (59%), involving 79 genes, map to different genomic clusters, indicating recent interchromosomal duplications (and/or tandem duplications followed by gross chromosomal rearrangements). In contrast, in mouse

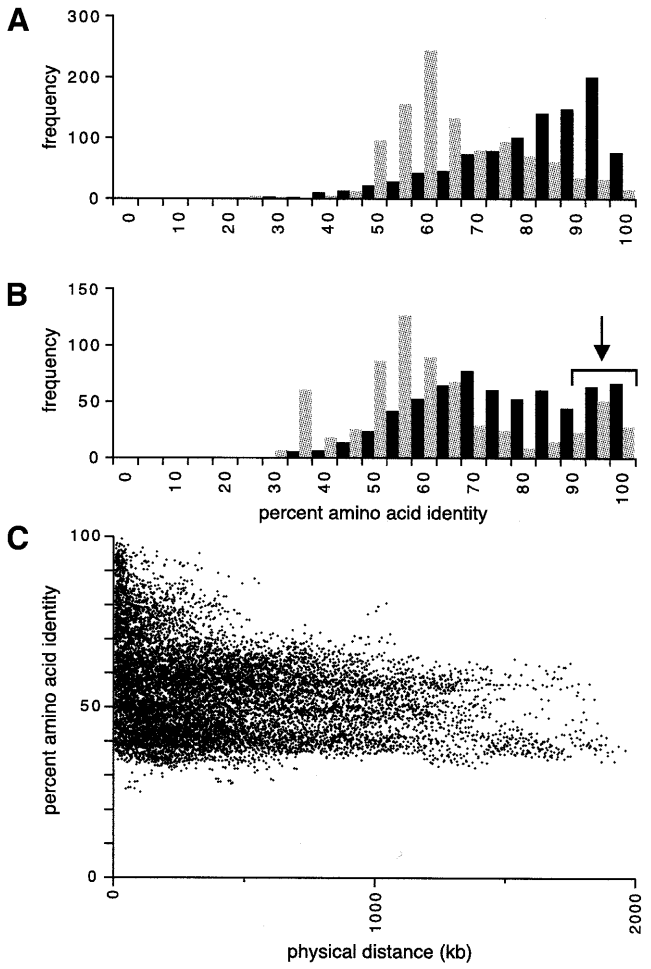

Figure 3. The OR gene family expands by local tandem duplications. Histograms show the distribution of amino acid identity of (**A**) each of the 1054 mouse genes in the full-length dataset with its most similar gene in another cluster (gray bars, average identity 64%) and its most similar gene in the same cluster (black bars, average identity 79%). (**B**) The same distributions for the 697 human genes from the HORDE dataset (4) with sequence length of ≥250 amino acids. Average identities of the best match in the same and different human clusters are 73 and 59%, respectively. The arrow marks a peak representing human genes, including the *OR7E* genes, which have recently undergone interchromosomal duplications. (**C**) Scatterplot comparing percent amino acid identity between pairs of intact sequences in the same cluster (*y*-axis) compared to their physical distance (*x*-axis).

only 33/207 (16%) of the gene pairs of >90% identity are in different clusters, indicating that most recent duplications in mouse are local in nature. Of the 79 dispersed human genes, 40 are members of the very large *OR7E* subfamily of pseudogenes, which has recently scattered to at least 35 places around the genome (4). Many of the other dispersed genes are in subfamilies *OR4F* and *OR4G*, representing OR genes in a multicopy subtelomeric sequence block (27) or in subfamily *OR4K*, representing genes found near the centromeres of several, predominantly acrocentric, chromosomes (see below and Fig. 5).

On a smaller scale, the percent identity of sequence pairs is weakly and inversely correlated to their physical separation (Fig. 3C). Within clusters, neighboring genes are also often in the same transcriptional orientation. Of the 1225 neighbor pairs in our comprehensive dataset, 850 (69%) are in the same orientation. This percentage is significantly more than the 50% expected if assortment was random and indicates that the tandem duplications are not generally associated with inversions.

Mouse locations are syntenic to human OR loci

Phylogenetic trees constructed using all full-length mouse and human OR sequences show that most major clades contain both mouse and human sequences (Fig. 6). This pattern suggests that most OR subfamilies were present in the common ancestor. We could identify orthologous locations in the human genome for 27 of the mouse OR-containing genomic locations, which together contain 1170 (92%) of the 1267 mapped OR genes (Fig. 5). Thus, most OR clusters were present when the primate and rodent lineages diverged and still exist now. The chromosomal locations of most pairs of orthologous clusters correspond to known syntenic blocks (http://www.ncbi.nlm.nih.gov/Homology/index.html) and include several previously described orthologous OR clusters (13–17). Figure 4 illustrates that mouse and human genes from two pairs of locations that we identify as orthologous indeed belong, with few exceptions, to the same major phylogenetic clades. Some matches between mouse and human gene clusters (Fig. 5, labeled in red) are not part of known syntenic relationships. In human, three groups of such clusters represent the genes subject to interchromosomal duplications (see above), both interstitially and in subtelomeric and pericentromeric regions. In mouse, two groups of genes appear to have spread to multiple chromosomes, although not near telomeres or centromeres.

Of the 1054 mouse genes in the full-length dataset, 836 (79%) have a human match of ≥70% nucleotide identity over ≥200 bp, indicating that a potential ortholog can be found. However, several genes may share the same ortholog (below and Fig. 5C) due to the expansion of many clusters by local duplications in both the mouse and human genomes. This phenomenon is particularly common in mouse, and two striking examples are shown in Figure 5A. A more detailed examination of mouse–human orthologous relationships (e.g. Fig. 5C) reveals many changes in both species since the primate and rodent lineages diverged, with the result that few genes have a single ortholog. Most mouse genes in the full-length dataset (809/1054 or 77%) have a closer relative in mouse than in human. Similarly, most human genes (548/906 or 60%) have a closer relative in human than in mouse. This observation is supported by a phylogenetic tree (Fig. 6), which shows many groups of mouse or human sequences that are more similar to one another than to any sequence from the other species. These species-specific sequence groups could arise from duplication and/or gene conversion since primate–rodent divergence or from loss of the orthologous gene(s) (loss from the genome or because the datasets are incomplete). Most of the 809 mouse–mouse best matches arose since primate–rodent divergence, since their level of identity is greater than is typical for orthologous genes of this family (Fig. 7).

DISCUSSION

We report here the results of a comprehensive analysis of orthologous and syntenic relationships of mouse and human

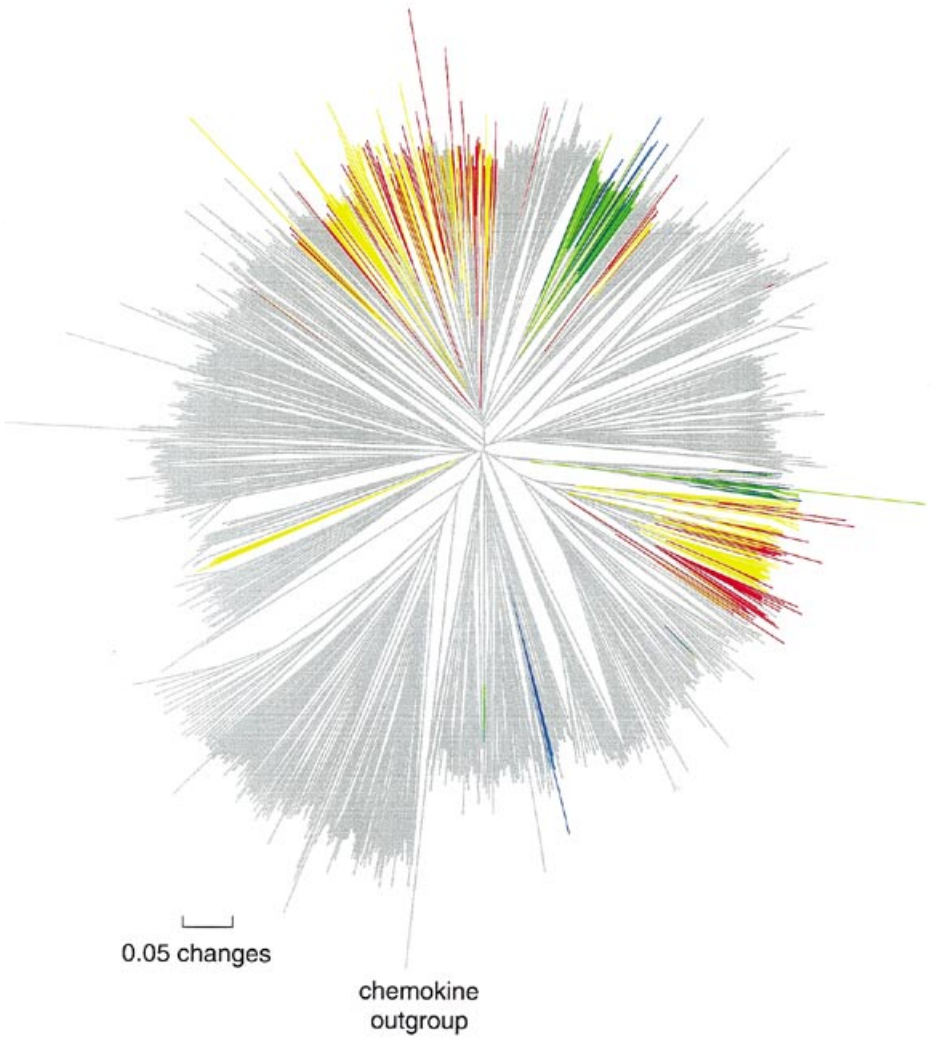

Figure 4. A phylogenetic tree highlighting the evolutionary relationships between OR genes at two pairs of orthologous locations in the mouse and human genomes. All mouse genes with full-length sequences (1045 sequences), all human genes from the HORDE dataset with ≥250 predicted amino acids (693 genes) and a chemokine receptor as an outgroup were aligned. A tree was drawn using the neighbor-joining algorithm of PAUP. The scale shows the branch length representing 0.05 changes per amino acid position. Sequences mapping to mouse chromosome 9A5 are colored green and sequences from the orthologous human location, chromosome 11q24.2, are in blue. Mouse chromosome 2E1 sequences are in yellow, and sequences at the orthologous human location, chromosome 11p11–11q12, are in red. Sequences from these pairs of locations cluster together on the tree, showing that potential orthologs have been identified.

OR genes. This analysis reveals striking differences in the size, functional constraints and evolutionary processes that have acted on the OR family since primate–rodent divergence.

The availability of an almost complete genome sequence has allowed us to identify the sequences of approximately 1400 mouse OR genes. From this number, we estimate that there are at least 1500 OR genes in the mouse genome. This is ~50% more than previously predicted (3), and ~50% more than is found in the human genome (4).

Frequent local duplications of OR genes in the rodent lineage are responsible for much of the size difference between the mouse and human OR families. These events are apparent on phylogenetic trees (Figs 4 and 6) and by comparison of the maps of the two OR subgenomes (Fig. 5). Several instances of new mouse OR genes created by local duplications were also noted in previous studies that compared clusters in the mouse and human genomes (13–15,17). Deletions of OR genes in the human lineage (14) have further exacerbated the differences between the two families. However, genes from both species persist in most major phylogenetic clades. In addition, murine clusters containing 92% of mapped OR genes have orthologous OR loci in the human genome (Fig. 5), showing that the gross arrangement of these gene clusters was established before primate–rodent divergence.

Despite their greater number, mouse OR genes are found at markedly fewer locations than human ORs (46 versus 104) (4). Whereas most recent activity in both mouse and human OR families has involved local duplication and/or gene conversion events, a subset of human OR genes has undergone recent duplication to distant locations in the genome, accounting for most of the increased genomic dispersion of the human OR gene family. These interchromosomal duplications often involve large blocks of sequence and are apparent by both our comparative sequence analyses and FISH (7,8). We observe

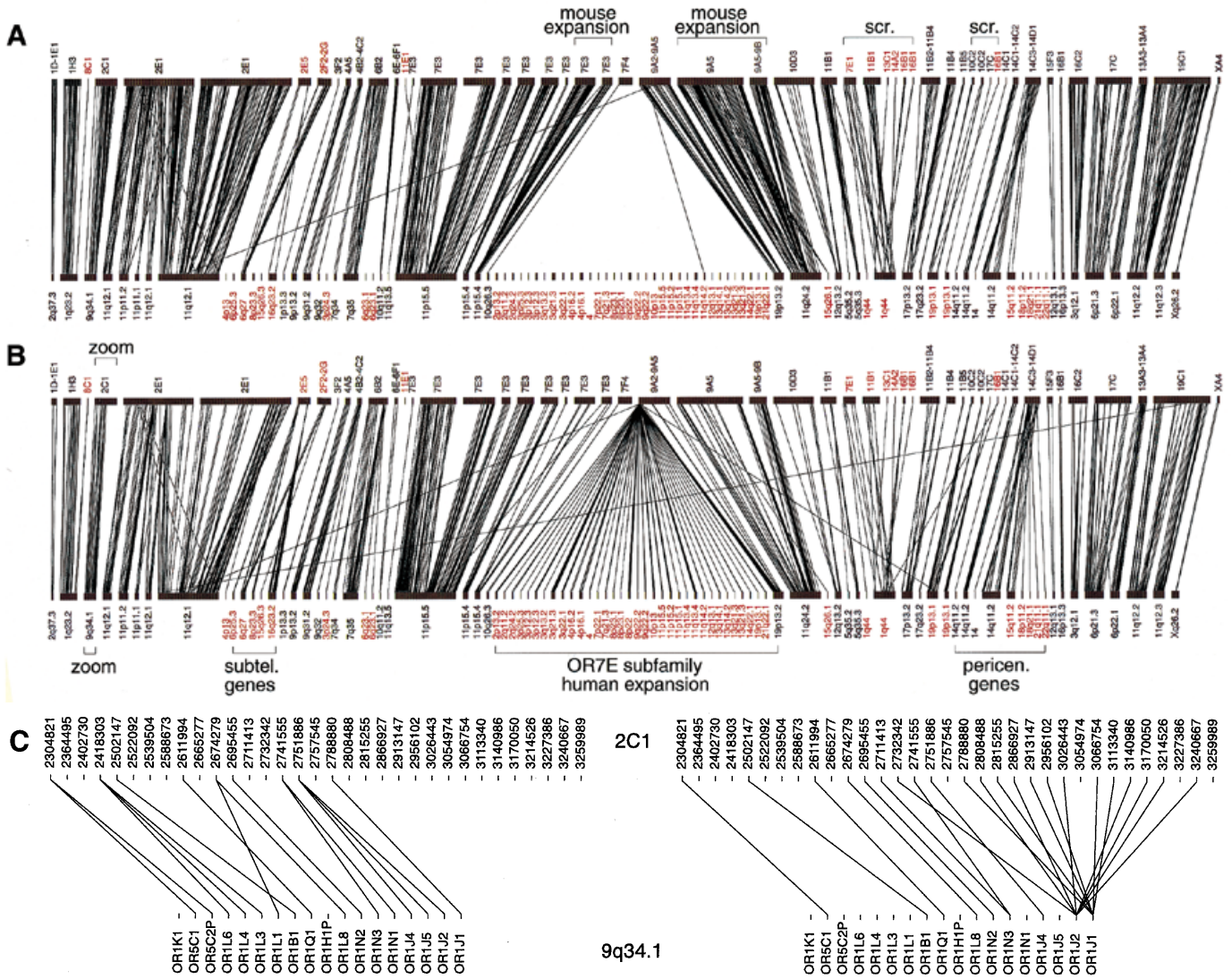

Figure 5. Orthologous relationships between mouse and human OR clusters. (**A**) A line is drawn from each mouse gene (upper row) to its most similar human gene (lower row) (considering only matches of ≥70% nucleotide identity over ≥200 bp). (**B**) A line is drawn from each human gene to its most similar mouse gene using the same identity threshold. Clusters where no gene has a cross-species match exceeding this threshold are not shown. Gene order is the same in (A) and (B) and was determined as follows: genes of each species are ordered within their clusters according to genomic positions in Celera mouse scaffolds (see Fig. 1 legend for mapping procedures) or the UCSC human genome assembly, but the clusters are rearranged to minimize the crossing of lines. Gene order within clusters was not changed, and clusters were not split. Gaps in the horizontal lines indicate breaks between clusters. (**C**) A selected part of the figure [labeled 'zoom' in (A)] at higher resolution, with mouse genes labeled according to their start position in scaffold GA_x5J8B7W4QPD. Human matches to each mouse gene are shown on the left, and mouse matches to each human gene on the right. In general, there is good colinearity between mouse genes and their best matching human genes, showing that most of the ancestral clusters have been maintained in both species. The figure also illustrates some local expansions [e.g. clusters labeled 'mouse expansion' in (A)]. Most pairs of similar OR gene clusters (labeled in black) fall into established syntenic chromosomal regions (http://www.ncbi.nlm.nih.gov/Homology/index.html), but some (labeled in red) do not. In human, there are three such groups of genes [labeled in (B)]: the *OR7E* subfamily, a group of genes in the subtelomeric regions of several chromosomes (27), and another group of genes mapping near the centromeres of several, primarily acrocentric, chromosomes. Conversely, two human chromosomal locations have matching genes at more than one mouse location, indicating scrambled synteny (labeled 'scr.'). The first human location is in the subtelomeric band 1q44, for which no syntenic mouse location was previously identified, perhaps because of the extreme scrambling of this region. The second location is at human chromosome 19p13.1, where OR genes were previously described as being near a syntenic breakpoint (17). Our analysis reveals an additional mouse locus with similar genes and suggests that the evolutionary history of this region is more complex than previously reported. Potential orthologs are not shown for all genes, in some cases because it was not possible to assign a map position to the best matching gene and in other cases because no gene matched above the identity threshold. Local order of some human genes may be wrong because many of the BAC sequences making up the genome assembly used for chromosomal localization (Materials and Methods) are unfinished. However, order and orientation of mouse genes is likely to be correct since Celera used a paired-end strategy to assemble the mouse genome sequence.

three groups of human OR genes involved in interchromosomal duplications (Fig. 5): the *OR7E* genes (4), the *OR4F* and *OR4G* genes located in subtelomeric regions (8,27) and a previously undescribed group, the *OR4K* genes, in the pericentromeric regions of several, predominantly acrocentric, chromosomes. Duplication of the *OR7E* genes to at least 35 locations is

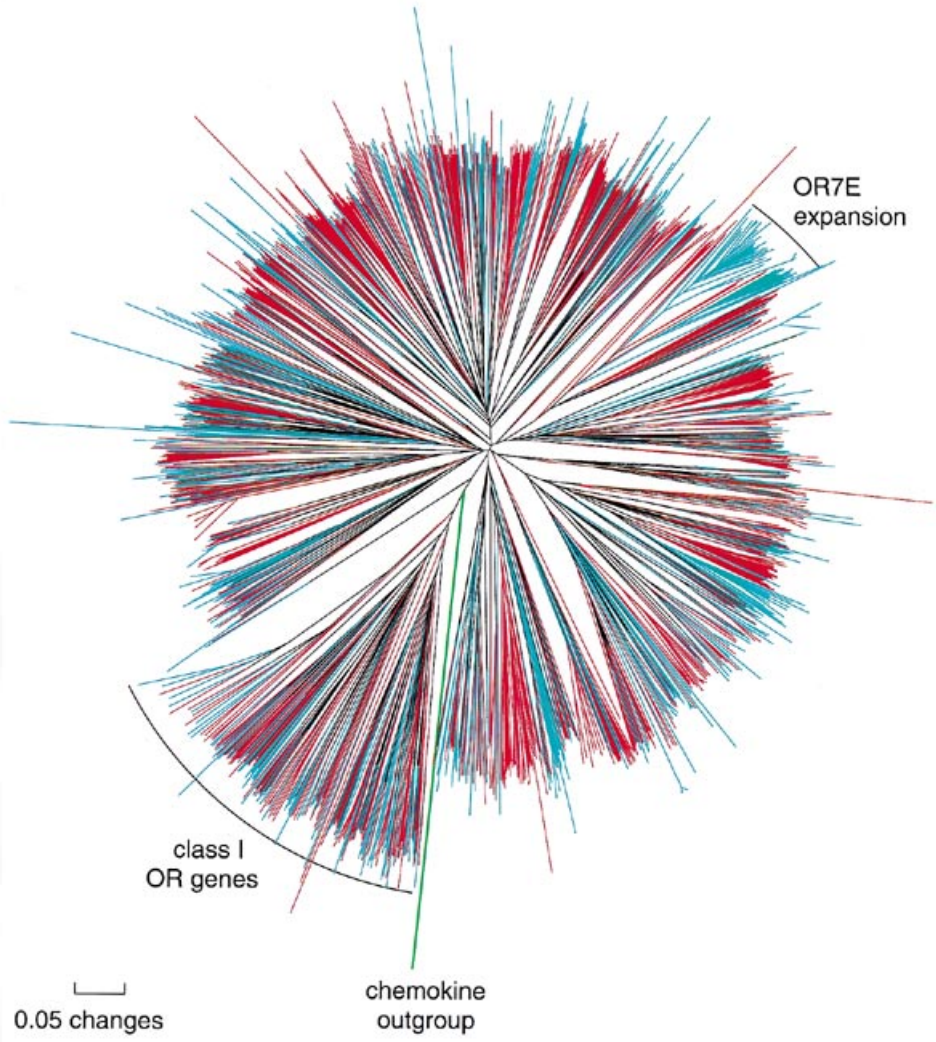

Figure 6. A phylogenetic tree showing the evolutionary relationships between all full-length mouse and human OR genes. The tree in Figure 4 is recolored with all mouse sequences shown in red and all human sequences shown in blue. Black parts of the tree indicate branches with both mouse and human 'descendants'. Examination of the longest branches reveals that they are pseudogenes, and are therefore released from selective pressure and diverge at a faster rate than intact genes.

especially surprising given that they all are pseudogenes (4); their duplication cannot be driven by selective pressure to increase the functional OR gene repertoire. In contrast, some of the subtelomeric and pericentromeric OR genes appear to be intact, and at least one is transcribed (28). The unusual evolutionary dynamics of these regions (27,29) may contribute to functional diversity in the human OR gene family.

The recent changes in the human and mouse OR families are reminiscent of the pattern of evolutionary changes observed for nematode chemosensory receptor genes (30,31) and other gene families for which sequence diversity is important, such as the eosinophil-associated RNase (32), MHC and immunoglobulin gene families (33). The evolution of these gene families is consistent with the gene 'birth-and-death' model, where new gene family members have arisen by gene duplication, followed by divergence and maintenance of some duplicate genes, and deletion or accumulation of mutations in other genes (33). In this model, the balance between rates of duplication and loss determines the size and pseudogene content of the gene family. Both the mouse and human families show a high rate of gene birth, as evidenced by the fact that over half of all genes in both species match another gene within the same genome better than one in the genome of the other species. Both species are losing genes, although far fewer pseudogenes are found in mouse than in human.

Approximately 20% of mouse OR genes are pseudogenes. Our estimate of pseudogene content is higher than was estimated previously from analysis of 33 genes (22). However, our sequence data-mining strategy enabled us to identify 134 pseudogenes (50% of the total) that are interrupted by interspersed repeat sequences. These pseudogenes would not have been amplified under the degenerate primer-based strategy used previously. The human OR gene family contains a much greater proportion of pseudogenes (63%) (4). This marked difference suggests a greater selective pressure in mouse to maintain a large functional OR repertoire, but may also be partly due to a faster elimination of pseudogenes from the mouse than the human genome (34).

The differences in family size and pseudogene fraction mean that the functional OR repertoire of mouse is more than three

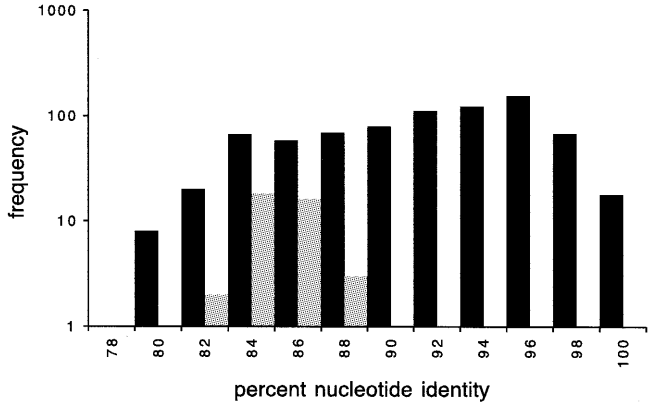

Figure 7. There have been many OR duplications since primate–rodent divergence. Distribution of percent nucleotide identities of putative mouse–human orthologous pairs (gray bars) and of 809 pairs of mouse genes and their best matches in the combined mouse–human dataset, where the best match was another mouse sequence (black bars). The average nucleotide identity of 41 orthologous gene pairs (Materials and Methods) is 85% (SD = 2%, range = 81–89%). Most of the mouse–mouse best matches are more similar than the orthologous pairs and are therefore likely to represent the products of duplication or gene conversion events since divergence of the primate and rodent lineages.

times larger than that of human (1180 versus 350 intact genes). The smaller repertoire is consistent with the observation that humans have a poor sense of smell compared to other mammals (1). Humans' decreased dependence on olfaction for survival, compensated for by an increased reliance on vision and hearing, would result in lower selective pressure to maintain or expand olfactory capabilities.

Our analysis of proteins encoded by intact mouse OR genes shows that they are more conserved than human ORs. Two possible explanations for the increased diversity in the human family are (i) positive selection to provide a diverse repertoire of odorant binding receptors and (ii) lower selective constraints on protein sequence in the human OR family. If the increased diversity was due to positive selection, one would expect most of the increased diversity to be in the variable regions thought to be important in ligand binding. However, we observe that the highly conserved parts of the protein thought to be important for functions common to all ORs are less conserved in the human family than the mouse family (e.g. Fig. 2). Loss of these conserved residues suggests that some apparently intact human ORs may not encode functional proteins.

Although a single clear ortholog can be identified for some OR genes (13,16), most genes have more than one 'ortholog' due to numerous changes in the mouse and human families since rodent–primate divergence. Local and interchromosomal duplications and/or gene-conversion events obscure many of the relationships between human and mouse genes that were once functional orthologs. Odorant ligands have been identified for a small number of rodent OR genes (35,36), and correlation with OR protein sequence could clarify structure–function relationships. Our analyses show that caution should be exercised when inferring receptor–ligand relationships across species, especially since even slight changes in receptor sequence can change the ligand that elicits the largest response (36).

OR genes are subject to a remarkable, but as yet undiscovered transcriptional control mechanism. Each OR gene is expressed in only one of four physical zones of the olfactory epithelium (37), and each olfactory neuron within a zone expresses only one allele (38) of a single OR gene (35,40). It is not known whether OR genes must be clustered in the genome for correct expression, or whether this arrangement exists simply because the gene family has expanded by tandem duplications. Control of expression may operate at the level of individual genes (via transcription factors or recombination), at the level of OR gene clusters (via a locus control mechanism or regulation of chromatin structure) or by stochastic mechanisms (37,38). The genomic context of expressed OR genes, as well as comparisons between the orthologous and paralogous OR genes and clusters identified by our study, will help elucidate these transcriptional control mechanism(s). Our comparative analysis of the mouse and human OR gene families will be useful in the study of this and other functional and evolutionary aspects of mammalian olfaction.

MATERIALS AND METHODS

Identification of BACs containing OR genes

Clones containing OR genes were identified by low-stringency hybridization with a probe generated by degenerate PCR of genomic DNA. Degenerate primers used for PCR matched conserved regions (transmembrane domains 2, 3, 6 and 7) of the gene family. One novel primer, TM3deg1 (5′-CAIA(C/T)IGCIAC(G/A)(A/T)AIC(T/G)(G/A)TC(G/A)TA-3′) was designed and other primer sequences were as described previously: OR5B, OR3B (41); P24, P28 (37); and P26 and P27 (35). Various primer combinations (OR5B/OR3B and P24/P28 with annealing temperature 40°C; TM3deg1/P28 and P26/P27 at 45°C) were used to amplify segments of mouse genomic DNA, as no single set of primers was expected to identify all OR genes. These low annealing temperatures were empirically determined to generate OR-specific probes when tested on Southern blots of BACs with known OR gene content. An initial denaturing step of 94°C for 5 min was followed by 35 cycles of 94°C for 30 s, annealing at temperatures stated above for 1 min, and extension at 72°C for 1 min, with a final extension at 72°C for 10 min. PCR products were labeled by inclusion of digoxigenin-11-dUTP (Roche Molecular Biochemicals) in the reaction and hybridized to nylon filters on which 158 900 recombinant clones (an estimated 11.2-fold genome coverage) from a mouse BAC library (RPCI-23) had been arrayed (42). A probe generated from human genomic DNA using the same PCR method was hybridized to filters containing 109 657 clones (6.7× coverage) from a human BAC library (RPCI-11) (43). This strategy anticipates that some OR genes may not be sufficiently similar to the rest of the gene family for both PCR primers to bind, but would be similar enough in the intervening sequence to be detectable by low-stringency hybridization. Filters were therefore hybridized and washed at low stringency (hybridization at 30°C in 5× SSC with 50% formamide; final wash in 2× SSC, 0.1% SDS at 65°C), and detected using chemiluminescence according to protocols recommended by Roche. To ensure high sensitivity, we chose BACs with both strong and weak hybridization signals. Recognizing that this approach could give false positives, we used PCR to confirm the presence of OR genes in 94 of the BACs, testing them with seven degenerate primer combinations

(OR5B/OR3B, P24/P28, P24/OR3B, TM3deg1/OR3B, TM3deg1/P28, OR5B/P28, P26/P27). We required that at least one primer pair give a product of the expected size. We obtained end sequences from OR-containing BAC clones as described previously (44) and from a publicly available resource (45).

Chromosomal localization of BAC clones by FISH

BACs were hybridized to mouse mitotic cells fixed to slides using procedures for FISH detailed elsewhere (46). Briefly, mouse metaphase spreads were prepared from spleen cell suspensions after lysis of red blood cells. Splenocytes were cultured for 48 h in lipopolysaccharide to stimulate cell cycling, arrested in mitosis by incubation in colcemid for 10 min prior to harvest, and fixed to glass slides using conventional cytogenetic methods. All BACs were streaked to obtain single colonies, and DNA was prepared using an Autogen 740 robot. BAC DNA was biotinylated by nick translation, and 200 ng was hybridized to chromosomes at 37°C in 50% formamide/2× SSC/10% dextran sulfate in the presence of 10 μg mouse Cot1 DNA, which suppresses labeling of interspersed repetitive elements. After washing in 50% formamide/2× SSC at 42°C, hybridization sites were labeled with avidin-FITC, the cells were washed, and they were then counterstained with DAPI applied in an antifade solution. Images were collected using a Zeiss Axiophot microscope equipped with ChromaTechnology spectral filters, a Photometrics Quantix cooled CCD camera, and IpLab Spectrum software. If a clone gave multiple FISH signals, BACs were streaked to obtain single colonies a second time, in order to exclude the possibility that the multiplicity of signals was due to mixed clones in the probe. From the 2471 OR-positive BACs, 130 BACs were chosen randomly and additional BACs were chosen based on their position in the BC Cancer Agency Genome Sequence Center's (BCGSC) physical map (http://www.bcgsc.bc.ca/). We FISH-mapped clones from any contigs that contained at least two OR hybridization-positive BAC clones, but did not contain any of the 130 randomly chosen clones. We also used the BCGSC contigs to choose BACs overlapping clones that gave multiple FISH signals, and at chromosomal locations where only one randomly chosen BAC had been mapped. When determining the number of OR-containing genomic loci, we counted only locations confirmed by having signals from at least two OR-containing clones.

Sequence database mining

A local database of OR protein sequences was compiled by downloading from GenBank any sequences annotated with the keywords 'olfactory receptor' or 'odorant receptor'. Some lamprey OR sequences were removed, because their closest mammalian homologs were serotonin rather than ORs (47). Other non-OR proteins were used as outgroups, including taste receptors, vomeronasal receptors, adrenergic receptors, melanocortin receptors and serotonin receptors. A similar set of mouse OR nucleotide sequences was downloaded from GenBank (234 sequences). This set was reduced to a non-redundant set of 155 sequences by taking only one representative of groups of sequences showing ≥97% sequence identity.

Celera's mouse genome assembly (http://www.celera.com/) was built from shotgun sequence reads representing a 5.25-fold coverage of the genome. At the time of our analysis (June 2001) it consisted of 19 778 'scaffold' sequences: sequences within which the order and orientation of the sequence should be correct, but containing gaps whose size can be estimated with reasonable accuracy. Scaffold sequences were searched using gapped tblastn (48) with 34 previously identified OR protein sequence queries, chosen based on phylogenetic diversity as assessed by preliminary sequence analysis (L.Linardopoulou and R.Lane, unpublished data) and by human OR gene classification (49), using one non-pseudogene member of each human OR 'family' where possible. The query set consisted of HORDE genes *OR1D5*, *OR2F1*, *OR3A3*, *OR4F5*, *OR5M8*, *OR6T1*, *OR7D2*, *OR8H2*, *OR9A4*, *OR10J1*, *OR11A2*, *OR12D3*, *OR13C5*, *OR51H1*, *OR52A1*, *OR55C1P* and *OR56A1* (4), genes from the mouse *P2* cluster, *I7*, *M50*, *B1*, *B2*, *B5*, *P2*, *P3* and *P4* (14), four subtelomeric human OR genes, *OR-7501A*, *OR-7501B* and *OR-7501C* (8) and *OR4F3*, as well as miscellaneous other genes; *C3* (36), *HSHTPRH06* (50), *K18* (37), *OR11-8c* (51) and an anonymous gene with GenPept accession no. AAC18915. Perl scripts were written to identify all genomic locations in scaffolds where an E score $\leq 10^{-5}$ was obtained with any of the query sequences and to extract these sequences with 1 kb of additional sequence on each side.

We used a modification of the method of Glusman *et al.* (4) to predict the OR protein sequence of each gene. Each potential OR gene and its flanking sequence was first screened for repeats using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and the Repbase database (52). Sequences were then compared to a local database of full-length OR protein sequences using fastx33 (53) to identify the best reading frame, allowing for frameshifts in the sequence. Sequences were then extended outward codon by codon to try to find suitable start and end codons.

After identification of potential OR genes (1986 sequences) and prediction of protein sequence, the following filters were applied to the data. (i) Pairs of OR gene fragments close in the genome and interrupted by repeat sequences, but appearing to be two halves of the same gene, were combined into one sequence (14 pairs). (ii) Sequences matching non-OR G-protein-coupled receptors better than ORs were eliminated (95 sequences). (iii) Sequences whose original blast hit was very weak and did not match OR genes by fastx33 comparison were eliminated (46 sequences). (iv) When only partial data was available (e.g. OR genes abutting an end or gap in Celera's scaffold sequence), we eliminated sequences with a match of ≥97% nucleotide identity over ≥200 bp to a sequence in the full-length data set (Results), reasoning that these sequences come from the same gene but, for some reason, were not properly assembled (184 sequences). (v) Apparent pseudogenes were required to match another OR gene (a previously identified OR or one of the 866 intact OR genes identified here) with ≥40% amino acid identity over 100 residues or ≥50% identity if between 25 and 99 amino acids. These criteria were chosen because all of the intact, full-length OR sequences we found matched a previously identified OR with ≥40% identity. These criteria are similar to those used by Glusman *et al.* (4) in their evaluation of the human OR family. Filter V resulted in the elimination of 178 sequences. Additional redundancy among the 262 partial sequences was eliminated by taking only one representative of each group of sequences sharing ≥97% sequence identity, leaving 187 unique sequences.

Two human datasets were analyzed. The sequences presented by Zozulya *et al.* (5) were used to determine the degree of conservation of intact human OR proteins. Sequences were obtained from HORDE (http://bioinformatics.weizmann.ac.il/HORDE/) for other analyses.

Chromosomal localization of sequences

Two sources of chromosomal localization information were available for the OR-containing scaffold sequences. FISH-mapped BACs were cross-referenced to the scaffold sequences when one or both of their end sequences matched the scaffold sequence with ≥95% sequence identity over three-quarters of their length. Unmasked end sequences were used for one-third of the sequences when less than 50 bp of unique sequence remained after repeat-masking. Matches were rejected if more than one genomic region matched the BAC end sequence at this level of identity. These mapping data were supplemented by chromosomal localization data made available by Celera. Celera has mapped scaffolds based on the linkage or radiation-hybrid map position of any sequence tagged sites matching scaffold sequences. For the small number of unmapped OR-containing scaffolds, we used matching end sequences to choose an additional 36 BACs to FISH, and thus localized another 17 scaffolds. Remaining unmapped scaffolds were small and had no matching BAC end sequences in GenBank (July 2001) or in our own BAC end sequence database.

Although HORDE supplies a chromosomal location for many human OR sequences, we updated and refined these positions by comparing each to the December 12, 2000 version of the UCSC genome assembly (http://genome.ucsc.edu/). We required ≥99% nucleotide match over ≥50 bp to assign a map position.

Sequence analysis

An initial sequence alignment was obtained using CLUSTALW (54) and edited by hand. PAUP v4.0b6 (Version 4, Sinauer Associates, Sunderland, MA) was used to determine protein divergences and to generate a phylogenetic tree using the neighbor-joining method (gaps of more than one amino acid in size were coded as one gap plus missing data for the rest of the positions). Tree branches were colored using a custom perl script. Alignments of all mouse genes with the 906 human OR genes identified by Glusman *et al.* (4) showed that 23 of the human sequences were not alignable to OR protein sequences—in fact when they were used to search public sequence databases, OR genes were not among the best matching sequences. These sequences were therefore eliminated from subsequent analysis, along with any sequences that Glusman *et al.* (4) were unable to classify into an OR family according to their system of nomenclature. One sequence in the human data set (*OR1E7*) is of mouse origin and was also eliminated. *OR1E7* exactly matches mouse sequences in both the public and Celera databases over its entire length. Six mouse sequences were not easily alignable and were removed. Gapped blast v2.2.1 (48) was used for other large-scale comparisons. We chose 41 mouse–human orthologous gene pairs conservatively; we used only gene pairs where neither gene was a pseudogene, and where there appear to have been no duplications in either species since primate–rodent divergence (i.e. the mouse gene was the best matching mouse gene of only one human gene, and this human gene was the best match of the same mouse gene and no other). The information content of protein sequence alignments was determined using alpro (26). For information content analysis, mouse and human alignments contained only equivalent residues; positions where most sequences had a gap in only one species were disregarded.

We developed a custom database using acedb (http://www.acedb.org/) and used it to store and cross-reference information about clones and sequences. Our website (http://www.fhcrc.org/labs/trask/OR) provides a database where mapping information and orthologous relationships can be queried. Under the terms of our agreement with Celera Genomics, we are able to provide on our website only the sequences of the 445 genes described in this paper that we have confirmed experimentally by isolating and sequencing cDNA clones (J.Young, J.Ross, E.Williams, T.Newman, L.Tonnes-Priddy, R.Lane and B.Trask, manuscript in preparation). Additional gene sequences will be released as we find more matching cDNAs, and genes in our database will be linked to any publicly available matching sequences. Mouse BAC-end sequences generated from OR-positive BACs have Genbank accession nos BH405737–BH406512.

ACKNOWLEDGEMENTS

We thank Bob Lane, Tera Newman and Ger van den Engh for helpful discussions, Greg Mahairas and Steve Swartzell of the Institute for Systems Biology for BAC end sequencing, and Martha Ogilvie of Celera for assistance with batch BLAST searches. The data in this paper were generated in part through use of the Celera Discovery System™ and Celera Genomics' associated databases. This work was supported by NIH grant R01 DC04209.

REFERENCES

1. Stoddart,D.M. (1980) *The Ecology of Vertebrate Olfaction*. Chapman and Hall, London and New York.
2. Buck,L. and Axel,R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, **65**, 175–187.
3. Buck,L.B. (1992) The olfactory multigene family. *Curr. Opin. Genet. Dev.*, **2**, 467–473.
4. Glusman,G., Yanai,I., Rubin,I. and Lancet,D. (2001) The complete human olfactory subgenome. *Genome Res.*, **11**, 685–702.
5. Zozulya,S., Echeverri,F. and Nguyen,T. (2001) The human olfactory receptor repertoire. *Genome Biol.*, **2**, 1–12.
6. Rouquier,S., Taviaux,S., Trask,B.J., Brand-Arpon,V., van den Engh,G., Demaille,J. and Giorgi,D. (1998) Distribution of olfactory receptor genes in the human genome. *Nat. Genet.*, **18**, 243–250.
7. Trask,B.J., Massa,H., Brand-Arpon,V., Chan,K., Friedman,C., Nguyen,O.T., Eichler,E., van den Engh,G., Rouquier,S., Shizuya,H. *et al.* (1998) Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.*, **7**, 2007–2020.
8. Trask,B.J., Friedman,C., Martin-Gallardo,A., Rowen,L., Akinbami,C., Blankenship,J., Collins,C., Giorgi,D., Iadonato,S., Johnson,F. *et al.* (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.*, **7**, 13–26.
9. Wysocki,C.J. and Beauchamp,G.K. (1984) Ability to smell androstenone is genetically determined. *Proc. Natl Acad. Sci. USA*, **81**, 4899–4902.
10. Whissell-Buechy,D. and Amoore,J.E. (1973) Odour-blindness to musk: simple recessive inheritance. *Nature*, **242**, 271–273.
11. Sosinsky,A., Glusman,G. and Lancet,D. (2000) The genomic structure of human olfactory receptor genes. *Genomics*, **70**, 49–61.

12. Mombaerts,P. (1999) Molecular biology of odorant receptors in vertebrates. *Annu. Rev. Neurosci.*, **22**, 487–509.
13. Bulger,M., Bender,M.A., van Doorninck,J.H., Wertman,B., Farrell,C.M., Felsenfeld,G., Groudine,M. and Hardison,R. (2000) Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse β-globin gene clusters. *Proc. Natl Acad. Sci. USA*, **97**, 14560–14565.
14. Lane,R.P., Cutforth,T., Young,J., Athanasiou,M., Friedman,C., Rowen,L., Evans,G., Axel,R., Hood,L. and Trask,B.J. (2001) Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl Acad. Sci. USA*, **98**, 7390–7395.
15. Lapidot,M., Pilpel,Y., Gilad,Y., Falcovitz,A., Sharon,D., Haaf,T. and Lancet,D. (2001) Mouse-human orthology relationships in an olfactory receptor gene cluster. *Genomics*, **71**, 296–306.
16. Younger,R.M., Amadou,C., Bethel,G., Ehlers,A., Lindahl,K.F., Forbes,S., Horton,R., Milne,S., Mungall,A.J., Trowsdale,J. *et al.* (2001) Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res.*, **11**, 519–530.
17. Dehal,P., Predki,P., Olsen,A.S., Kobayashi,A., Folta,P., Lucas,S., Land,M., Terry,A., Ecale Zhou,C.L., Rash,S. *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science*, **293**, 104–111.
18. Sullivan,S.L., Adamson,M.C., Ressler,K.J., Kozak,C.A. and Buck,L.B. (1996) The chromosomal distribution of mouse odorant receptor genes. *Proc. Natl Acad. Sci. USA*, **93**, 884–888.
19. Fan,W., Liu,Y.C., Parimoo,S. and Weissman,S.M. (1995) Olfactory receptor-like genes are located in the human major histocompatibility complex. *Genomics*, **27**, 119–123.
20. Carver,E.A., Issel-Tarver,L., Rine,J., Olsen,A.S. and Stubbs,L. (1998) Location of mouse and human genes corresponding to conserved canine olfactory receptor gene subfamilies. *Mamm. Genome*, **9**, 349–354.
21. Strotmann,J., Hoppe,R., Conzelmann,S., Feinstein,P., Mombaerts,P. and Breer,H. (1999) Small subfamily of olfactory receptor genes: structural features, expression pattern and genomic organization. *Gene*, **236**, 281–291.
22. Rouquier,S., Blancher,A. and Giorgi,D. (2000) The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc. Natl Acad. Sci. USA*, **97**, 2870–2874.
23. Freitag,J., Ludwig,G., Andreini,I., Rossler,P. and Breer,H. (1998) Olfactory receptors in aquatic and terrestrial vertebrates. *J. Comp. Physiol. [A]*, **183**, 635–650.
24. Hoppe,R., Weimer,M., Beck,A., Breer,H. and Strotmann,J. (2000) Sequence analyses of the olfactory receptor gene cluster mOR37 on mouse chromosome 4. *Genomics*, **66**, 284–295.
25. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
26. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
27. Mefford,H.C., Linardopoulou,E., Coil,D., van den Engh,G. and Trask,B.J. (2001) Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. *Hum. Mol. Genet.*, **10**, 2363–2372.
28. Linardopoulou,E., Mefford,H.C., Nguyen,O., Friedman,C., van den Engh,G., Farwell,D.G., Coltrera,M. and Trask,B.J. (2001) Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location. *Hum. Mol. Genet.*, **10**, 2373–2383.
29. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
30. Robertson,H.M. (1998) Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.*, **8**, 449–463.
31. Robertson,H.M. (2000) The large srh family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. *Genome Res.*, **10**, 192–203.
32. Zhang,J., Dyer,K.D. and Rosenberg,H.F. (2000) Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc. Natl Acad. Sci. USA*, **97**, 4701–4706.
33. Nei,M., Gu,X. and Sitnikova,T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl Acad. Sci. USA*, **94**, 7799–7806.
34. Graur,D., Shuali,Y. and Li,W.H. (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J. Mol. Evol.*, **28**, 279–285.
35. Malnic,B., Hirono,J., Sato,T. and Buck,L.B. (1999) Combinatorial receptor codes for odors. *Cell*, **96**, 713–723.
36. Krautwurst,D., Yau,K.W. and Reed,R.R. (1998) Identification of ligands for olfactory receptors by functional expression of a receptor library. *Cell*, **95**, 917–926.
37. Ressler,K.J., Sullivan,S.L. and Buck,L.B. (1993) A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell*, **73**, 597–609.
38. Chess,A., Simon,I., Cedar,H. and Axel,R. (1994) Allelic inactivation regulates olfactory receptor gene expression. *Cell*, **78**, 823–834.
39. Vassar,R., Ngai,J. and Axel,R. (1993) Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell*, **74**, 309–318.
40. Ngai,J., Chess,A., Dowling,M.M., Necles,N., Macagno,E.R. and Axel,R. (1993) Coding of olfactory information: topography of odorant receptor expression in the catfish olfactory epithelium. *Cell*, **72**, 667–680.
41. Ben-Arie,N., Lancet,D., Taylor,C., Khen,M., Walker,N., Ledbetter,D.H., Carrozzo,R., Patel,K., Sheer,D., Lehrach,H. *et al.* (1994) Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. *Hum. Mol. Genet.*, **3**, 229–235.
42. Osoegawa,K., Tateno,M., Woon,P.Y., Frengen,E., Mammoser,A.G., Catanese,J.J., Hayashizaki,Y. and de Jong,P.J. (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.*, **10**, 116–128.
43. Osoegawa,K., Mammoser,A.G., Wu,C., Frengen,E., Zeng,C., Catanese,J.J. and de Jong,P.J. (2001) A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.*, **11**, 483–496.
44. Mahairas,G.G., Wallace,J.C., Smith,K., Swartzell,S., Holzman,T., Keller,A., Shaker,R., Furlong,J., Young,J., Zhao,S. *et al.* (1999) Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. *Proc. Natl Acad. Sci. USA*, **96**, 9739–9744.
45. Zhao,S., Shatsman,S., Ayodeji,B., Geer,K., Tsegaye,G., Krol,M., Gebregeorgis,E., Shvartsbeyn,A., Russell,D., Overton,L. *et al.* (2001) Mouse BAC ends quality assessment and sequence analyses. *Genome Res.*, **11**, 1736–1745.
46. Trask,B. (1999) In Birren,B., Green,E.D., Hieter,P., Slapholz,S., Myers,R.M., Riethman,H. and Roskams,J. (eds), *Genome Analysis: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY. Vol. **4**, pp. 303-413.
47. Berghard,A. and Dryer,L. (1998) A novel family of ancient vertebrate odorant receptors. *J. Neurobiol.*, **37**, 383–392.
48. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
49. Glusman,G., Bahar,A., Sharon,D., Pilpel,Y., White,J. and Lancet,D. (2000) The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mamm. Genome*, **11**, 1016–1023.
50. Parmentier,M., Libert,F., Schurmans,S., Schiffmann,S., Lefort,A., Eggerickx,D., Ledent,C., Mollereau,C., Gerard,C., Perret,J. *et al.* (1992) Expression of members of the putative olfactory receptor gene family in mammalian germ cells. *Nature*, **355**, 453–455.
51. Buettner,J.A., Glusman,G., Ben-Arie,N., Ramos,P., Lancet,D. and Evans,G.A. (1998) Organization and evolution of olfactory receptor genes on human chromosome 11. *Genomics*, **53**, 56–68.
52. Jurka,J., Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
53. Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
54. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (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.