Mammalian Genome 11, 1016–1023 (2000).
DOI: 10.1007/s003350010196

Mammalian Genome
Incorporating Mouse Genome

The olfactory receptor gene superfamily: data mining, classification, and nomenclature

Gustavo Glusman,[1] Anita Bahar,[1] Dror Sharon,[1,*] Yitzhak Pilpel,[1,] Julia White,[2] Doron Lancet[1]**

[1]Department of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel
[2]MRC Human Biochemical Genetics Unit, University College London, 4 Stephenson Way. London NW1 2HE, United Kingdom

Received: / Accepted:

Abstract. The vertebrate olfactory receptor (OR) subgenome harbors the largest known gene family, which has been expanded by the need to provide recognition capacity for millions of potential odorants. We implemented an automated procedure to identify all OR coding regions from published sequences. This led us to the identification of 831 OR coding regions (including pseudogenes) from 24 vertebrate species. The resulting dataset was subjected to neighbor-joining phylogenetic analysis and classified into 32 distinct families, 14 of which include only genes from tetrapodan species (Class II ORs). We also report here the first identification of OR sequences from a marsupial (koala) and a monotreme (platypus). Analysis of these OR sequences suggests that the ancestral mammal had a small OR repertoire, which expanded independently in all three mammalian subclasses. Classification of "fish-like" (Class I) ORs indicates that some of these ancient ORs were maintained and even expanded in mammals.

A nomenclature system for the OR gene superfamily is proposed, based on a divergence evolutionary model. The nomenclature consists of the root symbol 'OR', followed by a family numeral, subfamily letter(s), and a numeral representing the individual gene within the subfamily. For example, OR3A1 is an OR gene of family 3, subfamily A, and OR7E12P is an OR pseudogene of family 7, subfamily E. The symbol is to be preceded by a species indicator. We have assigned the proposed nomenclature symbols for all 330 human OR genes in the database. A WWW tool for automated name assignment is provided.

Introduction

Olfactory receptors (ORs) are seven-transmembrane domain (7TM) proteins (Lancet and Pace 1987; Reed 1990; Buck and Axel 1991). Previous work (Lancet and Ben-Arie 1993) suggested their classification into at least eight families within the G protein-coupled receptor (GPCR) superfamily. OR genes are expressed with clonal and allelic specificity, with one specific OR gene being expressed in every olfactory cell (Lancet 1986; Chess et al. 1994). In contrast to the immunoglobulin system, where proteins specific for diverse antigenic ligands are generated by a complex system of somatic recombination and clonal selection, olfactory receptors are present in the genome in a large germ-line repertoire, estimated to consist of several hundred genes in mammalian species and about 100 genes in catfish [reviewed in (Mombaerts 1999)]. This suggests a large expansion of the OR repertoire in higher vertebrates. The ligand-binding phenomenology of ORs can be described by a probabilistic model (Lancet et al. 1993b), with many receptors binding many ligands with different affinities. Such combinatorial coding has been demonstrated (Malnic et al. 1999). In this context, the addition of novel receptors confers the advantage of broadening the ligand spectrum that can be recognized. Conversely, OR gene loss can cause specific anosmias, or reduced discriminating capability (Lancet et al. 1993a).

OR genes are intronless in their coding region (Buck and Axel 1991; Nef et al. 1992), but have a long intron splitting the 5′ untranslated region, as predicted by computer analysis of genomic sequence (Glusman et al. 1996) and confirmed by comparison of cDNA and genomic sequences (Asai et al. 1996) and transcription analysis (Qasba and Reed 1998; Walensky et al. 1998). OR genes have been found to be organized in the mammalian genome in many clusters (Ben-Arie et al. 1994; Griff and Reed 1995; Sullivan et al. 1996; Rouquier et al. 1998b). One of these clusters, fully sequenced by us (Glusman et al. 2000) on human Chromosome (Chr) 17 (17p13.3), includes 17 OR genes out of the expected several hundred in the human olfactory subgenome. The OR genes in this cluster belong to various families and subfamilies. Conversely, genes from the same family have been found in different clusters and on different chromosomes (Sullivan et al. 1996; Rouquier et al. 1998b), suggesting a complex history of gene and cluster duplications.

Prior to the present report, OR databases included several hundred annotated olfactory receptor genes from many species. Several methods have been used to assign "trivial" names to related sets of sequences, based on clone name (e.g., HGMP07E, R2C4), a cloning method or environment (e.g., HPFH1OR, HSOLFMF), a chromosomal location (e.g., OR17-2 or even OR912-95 for a group of chromosomes), a genome-wide sequential numbering with species assignment (e.g., OLFR89, ZF2A, SCor35), or an arbitrary designation (e.g., gen147). Moreover, various different roots have been used for denoting olfactory receptors, including OR, OLF, and OLFR. The consistent nomenclature system proposed here would thus be highly valuable for future research and inter-group communication in the olfactory receptor field.

Materials and methods

Cloning of monotreme and marsupial ORs. Platypus (*Ornithorhynchus anatinus*) and koala (*Phascolarctos cinereus*) DNA was kindly provided by Bronwyn Houlden (School of Biological Sciences University of New South Wales, Sydney, NSW 2052, Australia). The primers for PCR amplification were previously designed to amplify part (TM2-TM7) of the open reading frame of OR genes, based on the 5B and the 3B redundant primers (Ben-Arie et al. 1994). The PCR mixture contained a total volume of 25 μl containing 50 mM KCl, 10 mM Tris-HCl pH 8.3, 1.5 mM $MgCl_2$,

* Current address: Ocular Molecular Genetics Lab, Harvard Medical School, Massachusetts Eye and Ear Infirmary, 243 Charles St., Boston, MA 02114

** Current address: Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115

Submitted GenBank entries: AF262696–AF262732 and AF262950.

Correspondence to: D. Lancet; E-mail: Doron.Lancet@weizmann.ac.il

200 mM of each deoxyribonucleoside triphosphate (dNTP), 0.1 mM of each primer, 1 unit of *Taq* DNA polymerase (Boehringer Mannheim, Germany), and 50 ng of genomic DNA. PCR products were electrophorated in a 1% agarose gel to view their size, and then cloned into the CloneAMP pAMP1 system for rapid cloning kit (GibcoBRL). Plasmid DNA for sequencing was purified with the Wizard Plus SV Minipreps kit (Promega).

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. The primers used for sequencing of the cloned insert in the pAMP vector were as follows: 5′-AAGCTTGGATCCTCTAGAGC-3′ and 5′-CTGCAGGTACCGGTCCGG-3′. After base calling with the ABI Analysis Software (version 3.0), the analyzed data were edited using Sequencher 3.0 (GeneCodes Corporation, Ann Arbor, Mich.).

Estimation of repertoire size. The distribution of the number of monotreme (or marsupial) ORs that appeared n times as obtained experimentally was fitted to a set of binomial distributions with repertoire sizes ranging from 18 (or 20) to 500, assuming an equal probability for each gene to be cloned. The distributions resulting from each tested repertoire size were compared with the experimentally observed distribution, and a correlation coefficient was calculated for each.

Database search. tblastn (Altschul et al. 1997) was used to compare amino acid query sequences to the non-redundant version of GenBank 112, with a non-stringent expectation value cutoff of 1e-4. In addition, a set of 131 private sequences was used, including 94 human genomic sequences (Fuchs et al. in preparation) and the marsupial and monotreme sequences. For each database hit, a list of one or more distinct locations showing significant similarity to OR sequences was created; the relevant nucleotide sequence was extracted from the database and was further considered as a new OR gene candidate.

Composite conceptual translation. The nucleotide sequences of the OR candidates were translated conceptually with FASTY (Pearson et al. 1997) against a curated set of 96 OR protein sequences from human, dog, mouse, rat, chick, clawed frog, zebrafish, and catfish, available electronically. This translation method takes into account the possibility of frameshifts and reconstructs the original reading frame of the query, as compared with the most similar database hit. Since the latter frequently includes only part of the OR open reading frame, the translated region was extended in both directions (whenever possible) based on additional database hits. To do this, the query sequence was translated by using as template for translation database hits of lower similarity (higher expectation value, up to 1e-2) but more extensive than the initial, best hit. Any resulting translated segments that extend the original result to the 5′ or to the 3′ were suitably concatenated to build a composite translation product.

Range selection. Unfiltered blastp (Altschul et al. 1997) was used to compare each composite conceptual translation with the 216-amino acid-long segment of OR3A1 (Crowe et al. 1996; Glusman et al. 1996) spanning the region of transmembrane 2 to transmembrane 7, as defined by the OR5B and OR3B primers (Ben-Arie et al. 1994). This corresponds to positions 68–283 in the alignments of olfactory receptor genes in the G-protein Coupled Receptor Database, GPCRDb (http://swift.embl-heidelberg.de/7tm/). Resulting TM2-TM7 segments were used further only if they had a length of at least 100 amino acids.

Sequence collection ruleset. An iterative sequence search strategy was designed and implemented in Perl. This strategy was used to detect all OR-like sequences in the databases. Six initial amino acid query sequences were used: OR17-4, HSOLF3, OR17-40, HSOLF1, OR11-55, and the catfish ICTORDF, representing families 1, 2, 3, 5, 6 and Class I, respectively. Candidate OR-like sequences were translated conceptually, and the TM2-TM7 segment was recognized. Each new candidate sequence was compared by using blastp (Altschul et al. 1997) with the growing dataset of the previously detected OR sequences, with the following ruleset: if the new sequence is 100% identical to a previously detected sequence of the same species, it is considered redundant; if the level of identity is at least 60% (same OR subfamily), the new sequence is added to the dataset and not processed further; if the level of identity is at least 40% (same OR family), the new sequence is added to the dataset and treated as a new query automatically; if the new sequence is at least 40% identical to a previously identified outgroup (non-OR) sequence, it is added to the set of outgroup sequences; otherwise (less than 40% identity to any sequence) the new sequence is left for a human curator to classify (as a new OR family, and therefore as a new query, or as a new outgroup) based on additional information. After no more sequences were found, nrdb90 (Holm and Sander 1998) was used to optimally reduce redundancy to 99% level while keeping the longest possible sequences.

Multiple alignment and tree building. The OR amino acid sequences were aligned with ClustalW (Thompson et al. 1994; Higgins et al. 1996) by using standard parameters. ClustalW implements the NJ (neighbor-joining) method of Saitou and Nei (1987). The following human sequences were used as outgroups: the endothelial differentiation protein edg-1 (EDG1, locus HUMEDG), the MC2-melanocortin receptor (MC2R, locus HSACTHR), the A2a adenosine receptor (ADORA2A, locus HSA2AREC), and the β3 adrenergic receptor (ADRB3, locus HSB3A), which was used for rooting. Phylogenetic trees were depicted by using TreeView (Page 1996) on a Power Macintosh. Large (>500 taxa) trees were rooted by using retree from the Phylip package Version 3.57c (Felsenstein 1989).

Clustering of sequences into families and subfamilies. For each node N in the rooted NJ tree produced, the average distance AD is defined as:

$$\mathrm{AD}(N) \equiv \sum_{i,j} \mathrm{dist}(a_i,b_j);\ \forall a_i \in A,\ \forall b_j \in B \text{ and}$$

$$\mathrm{dist}(a,b) \equiv \frac{100 - \mathrm{PID}(a,b)}{100}$$

where A and B are the subtrees joined by node N, and PID(a,b) is the percentage identity between taxa a and b. Families and subfamilies are then defined as the largest subtrees rooted at nodes with AD $\leq$ 0.6 or 0.4, corresponding to at least 40% or 60% identity, respectively.

Results and Discussion

Sequence collection. An iterative, semi-automated data mining procedure was employed for retrieving all available OR sequences from the databases. The procedure required 34.5 h to complete, with up to five parallel processors on a Sun Enterprise 10000 computer, evidencing the magnitude of the gene superfamily under study. In total, 315 sequences were used as queries for blast searching, and 1822 resulting candidates were conceptually translated. Of these, 772 sequences were found by annotation or by additional comparisons to belong to other GPCR families, e.g., dopamine and serotonin receptors. Among the remaining sequences, 107 showed less than a minimal overlap with the sequence core between transmembrane helices 2 and 7 (TM2-7), and 112 were identical to other OR sequences of the same species. After eliminating all these 991 sequences, the resulting dataset included 831 OR protein sequences from 25 species (Table 1). Human intervention was required in 109 cases, 25 of which resulted in the recognition of new sets of OR genes. The facts that only about half of the candidate homologous sequences recognized prove to be OR genes (indicating low specificity) and that the detected non-OR sequences belong in many different GPCR families (not shown) suggest that the search procedure is very sensitive. Therefore, very few OR sequences are expected to have been missed. The high throughput genome sequences (HTGS) partition of GenBank was not searched, to avoid the inclusion in the phylogenetic analysis of low-quality sequence.

A conceptual translation method (based on FASTY) was used, which optimizes reading frame usage and reconstructs the original sequence, in case of frameshifted pseudogenes. This strategy was preferred, since for the purpose of classification it is important to

Table 1. Summary of olfactory receptor genes detected.

	Species (subtotal)	Entries	Redundant	Fraction
HS	human	391	45	47.05%
PT	chimp	23	2	2.77%
PN	bonobo	4		0.48%
GO	gorilla	25		3.01%
PP	orang	4		0.48%
HL	gibbon	2		0.24%
MA	macaque	14		1.68%
	Primates	*463*		*55.72%*
BT	cow	1		0.12%
SC	dolphin	17		2.05%
CF	dog	22		2.65%
SS	pig	20		2.41%
MM	mouse	106	2	12.76%
RN	rat	59	1	7.10%
	Eutheria	*688*		*82.79%*
PC	koala	20		2.41%
OA	platypus	16		1.93%
	Mammalia	*724*		*87.12%*
GG	chick	15		1.81%
	Amniota	*739*		*88.93%*
XL	xenopus	14		1.68%
NM	salamander	11		1.32%
RE	frog	5		0.60%
	Tetrapoda	*769*		*92.54%*
LC	coelacanth	8		0.96%
DR	zebrafish	25	1	3.01%
IC	catfish	9		1.08%
CA	goldfish	8		0.96%
	Vertebrata	*819*		*98.56%*
LF	lamprey	10		1.20%
	Craniata	*829*		*99.76%*
AM	honeybee	2		0.24%
	Totals	831	51	100.00%

Statistics of OR genes found, organized by species, and indicating the total number of entries, the number of redundant entries (to 99% identity), and their fraction of the total dataset. Italics indicate cumulative partial statistics for major phylogenetic groupings. Species codes used are: HS, *Homo sapiens;* PT, *Pan troglodytes;* PN, *Pan paniscus;* GO, *Gorilla gorilla;* PP, *Pongo pygmaeus;* HL, *Hylobates lar;* MA, *Macaca mulatta;* BT, *Bos taurus;* SC, *Stenella coeruleoalba;* CF, *Canis familiaris;* SS, *Sus scrofa;* MM, *Mus musculus;* RN, *Rattus norvegicus;* PC, *Phascolarctos cinereus;* OS, *Ornithorhynchus anatinus;* GG, *Gallus gallus;* XL, *Xenopus laevis;* NM, *Necturus maculosus;* RE, *Rana esculenta;* LC, *Latimeria chalumnae;* DR, *Danio rerio;* IC, *Ictalurus punctatus;* CA, *Carassius auratus;* LF, *Lampetra fluviatilis;* AM, *Apis mellifera.*

build an evolutionarily correct tree of the OR sequences and to avoid errors that could be introduced by "literal" translation. Since a large proportion of primate OR genes have been shown to be pseudogenes (Rouquier et al. 1998b, 2000; Sharon et al. 1999), such approach is particularly important for studying OR sequences. Indeed, a large proportion of the resulting sequences included a small number of frameshift corrections that allowed the reconstruction of sequences showing significant similarity to other OR genes throughout the resulting open reading frame. A similar procedure was used to detect and reconstruct known and additional pseudogenes of other gene families (G. Glusman, unpublished results), suggesting that the procedure may be of general use.

Many almost-identical OR sequences are present in the databases, in many cases derived from total genomic DNA or from flow-sorted chromosomes, and may be artefactual, representing the same locus with small sequence variations. Therefore, almost identical sequences were removed from the dataset to a 99% identity level, yielding an OR dataset of 780 sequences. This procedure removed mainly human entries (Table 1). Several cases exist of nearly identical sequences for which additional information (e.g., chromosomal location or mapping) indicates that they represent distinct genes. In these cases, manual analysis allowed us to maintain the sequences in the database.

Sequence classification. The 780 OR sequences were multiply aligned by using ClustalW. A phylogram was generated by using the neighbor-joining algorithm and was rooted with human β3 adrenergic receptor as outgroup. Additional non-OR GPCR sequences introduced into the analysis were found not to intermix with the OR genes, as expected. The clustering results (Fig. 1) show a major subdivision that includes OR of both Class I ("fish-like") and Class II (tetrapodan), as previously defined, based on a study in the amphibian *Xenopus laevis* (Freitag et al. 1995). The two classes were clearly separated in the present analysis, with 64 sequences in Class I and 704 sequences in Class II. Another major subdivision included six sequences from lamprey (*Lampetra fluviatilis*) reported as ORs (Berghard and Dryer 1998), which showed highest similarity to an adenosine receptor, which was previously used as an outgroup for comparing lamprey ORs to Class I and Class II ORs (Freitag et al. 1999). Interestingly, these lamprey sequences lack the MAYDRY motif typical of ORs (Lancet and Ben-Arie 1993), but most have instead an IAYDRY motif, which is more similar to the IAIDRY motif of the A2a adenosine receptors. This raises the interesting possibility that the I→Y mutation in this highly conserved motif happened at least twice independently during the very early evolution of the OR gene superfamily. This highly conserved, OR-specific tyrosine residue shows a significant bias in synonymous codon usage (Conticello et al. 2000), suggesting that its presence is important to OR function.

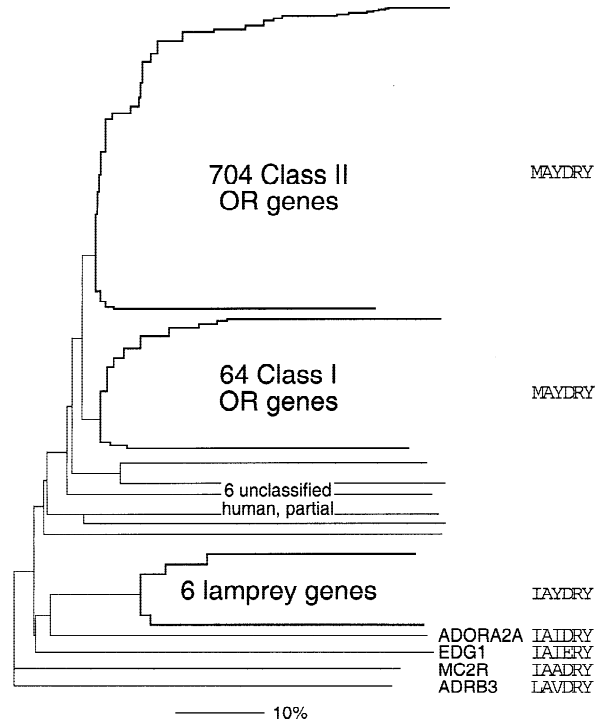

Fig. 1. Clustering results of the OR superfamily. The neighbor-joining tree was rooted by using ADRB3 (β3 adrenergic receptor). Additional outgroups are EDG1 (endothelial differentiation protein edg-1), MC2R (MC2-melanocortin receptor), and ADORA2A (A2a adenosine receptor). The right column shows the respective motif matching the OR-specific MAYDRY motif. The bar indicates 10% divergence along each branch.

Class I ORs. OR genes described to date in fish all belong to Class I, except for a few genes in the 'living fossil' coelacanth *Latimeria chalumnae* (Freitag et al. 1998). On the other hand, while in tetrapods a majority of the OR genes belong to Class II, many non-fish species (mammals, birds, and amphibians) were found to have Class I ORs as well. With average distance cutoffs corresponding to 40% and 60% protein sequence identity, our current phylogenetic analysis suggests that Class I ORs may be classified into at least 17 families (squares in Fig. 2). Interestingly, these families can be grouped further into five distinct "subclasses" that show a

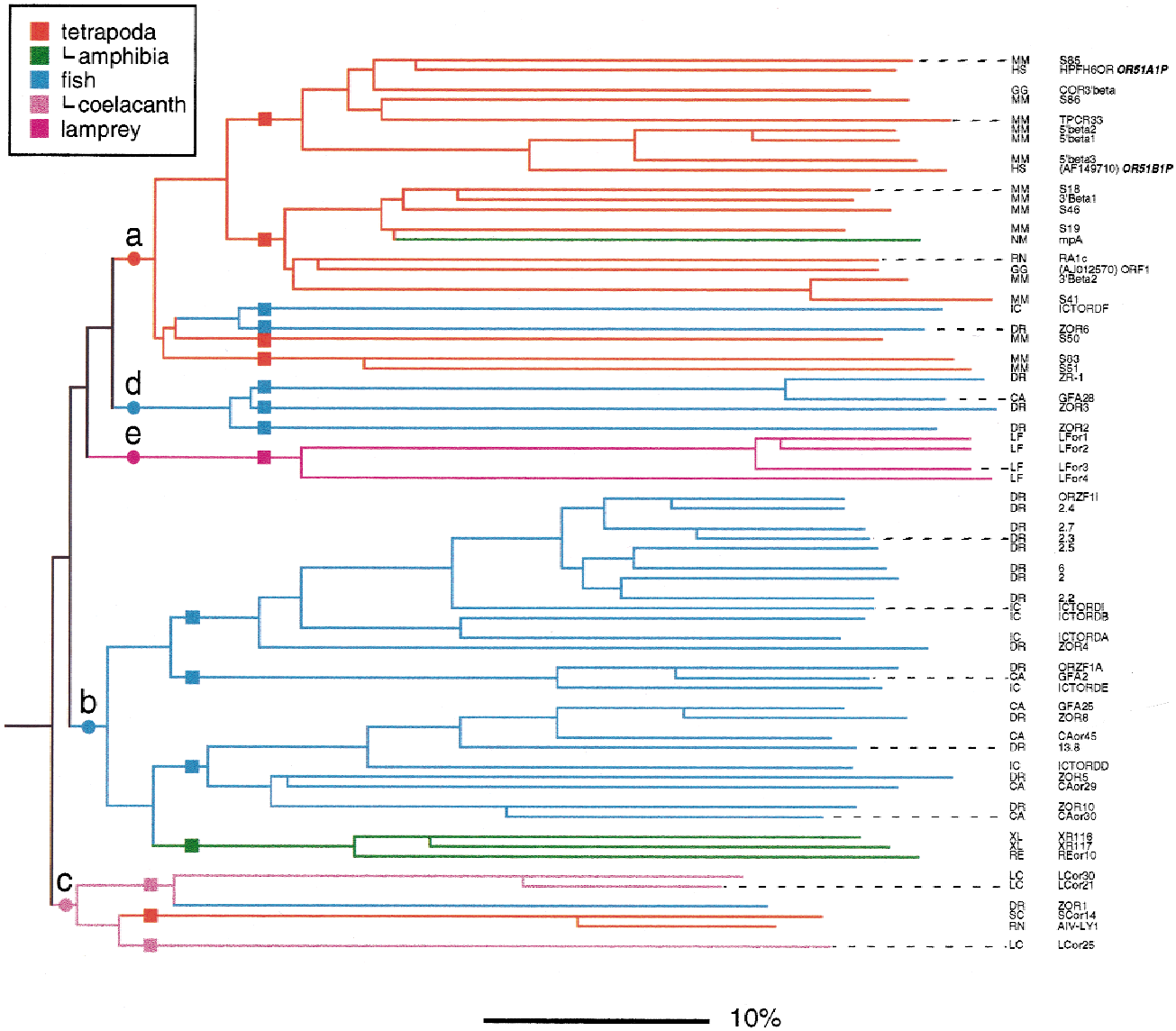

Fig. 2. Phylogenetic tree of the Class I OR genes detected. Circles indicate subclasses, and boxes indicate families. Tree elements are colored according to their most prominent taxonomical component. The columns on the right indicate species codes (as in Table 1) and trivial names. For human genes, a gene symbol is indicated. The bar indicates 10% divergence along each branch.

significant correlation to evolutionary classification (circles in Fig. 2). Subclass a includes mostly mammalian ORs, but also a few from bird, amphibians, and fish. The second large subclass b includes mostly fish ORs, but also a family of amphibian ORs. Subclass c includes coelacanth, mammalian, and fish representatives. Finally, two minor subclasses d and e include only fish and lamprey ORs, respectively.

While additional Class I families and subclasses may be uncovered in the future, it is apparent that the early vertebrate genome included a very small number of Class I precursor ORs. It can be hypothesized that the two successive tetraploidization events (Lundin 1993) could have raised the Class I OR count from an initial value of one or two genes to four to eight, currently represented by the subclasses. Of these, one subclass underwent significant expansion only in teleost fish, while another thrived in tetrapods.

In the human genome, Class I ORs are a minority. Ironically, one of the first ORs to be discovered was HPFH1OR, experimentally characterized as an enhancer causing hereditary persistence of fetal hemoglobin type 1 (HPFH1; Feingold and Forget 1989), and described as an apparent truncated GPCR, before the discovery of olfactory receptors (Buck and Axel 1991). Only 9 years later was it recognized (Buettner et al. 1998) as a rare human Class I OR. The same analysis also led to the identification of another human HPFH enhancer, HPFH6OR (Kosteas et al. 1997) as a Class I OR. Several additional examples of mammalian Class I ORs were subsequently discovered (Raming et al. 1998; Bulger et al. 1999). A more recently published corrected sequence for HPFH1OR showed higher resemblance to Class I ORs and was shown to be expressed in erythrocytes (Feingold et al. 1999). The surprising conservation and expansion of Class I ORs in terrestrial animals suggests that they may have attained new functions, which need not be related to olfaction.

Class II ORs. The Class II sequences were grouped into 14 families, each subdivided into several subfamilies. There is considerable asymmetry in the size of the different families: the families

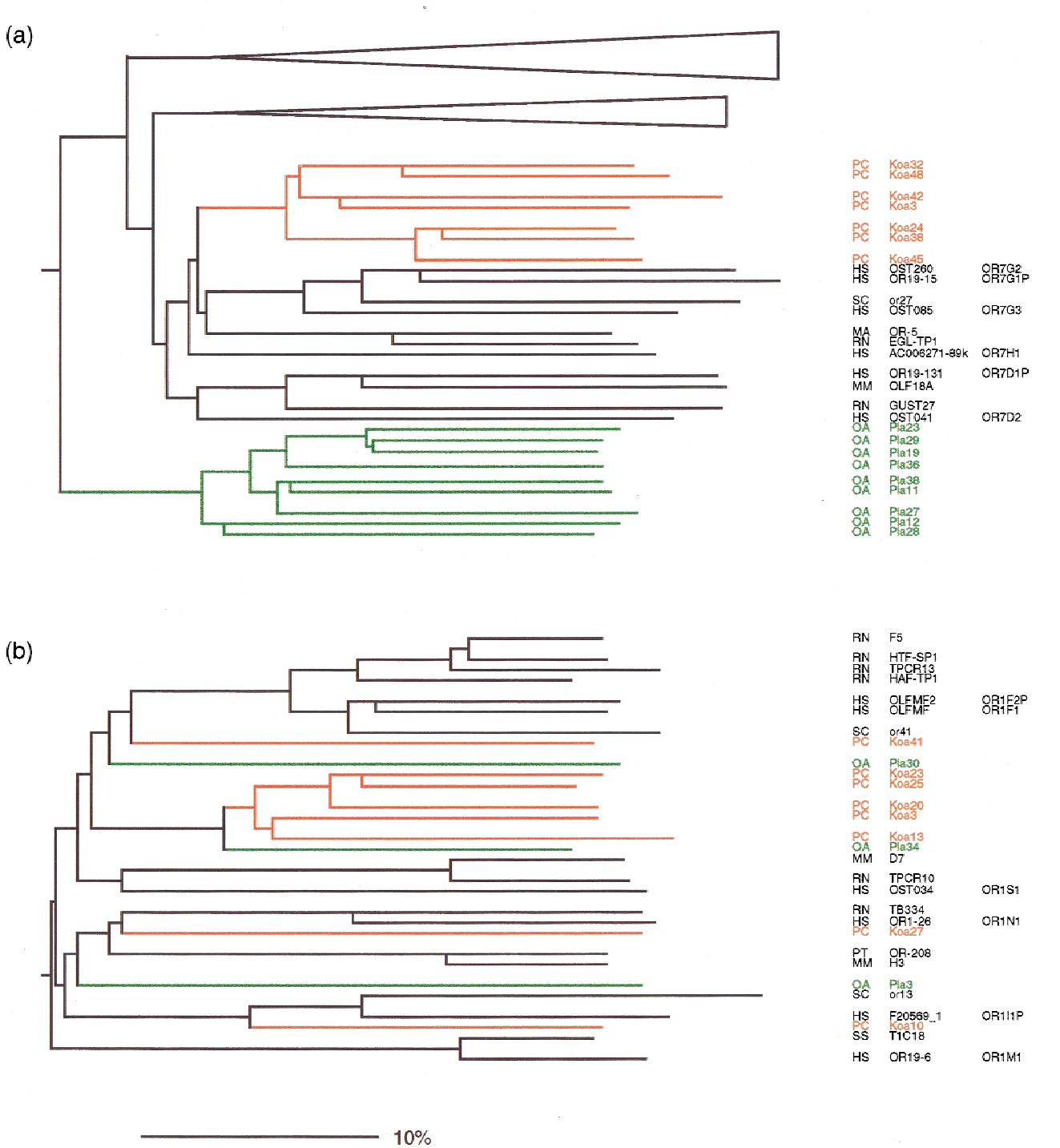

Fig. 3. Partial phylogenetic trees within family 1 (redundancy reduced to less than 90% identity), including most monotreme and marsupial ORs discovered. The columns on the right indicate species codes (as in Table 1) and trivial names. For human genes, a gene symbol is indicated. The bar indicates 10% divergence along each branch. **a)** The group of subfamilies defined as family 7. All platypus ORs are clustered, as well as all koala ORs. Triangles on top indicate subfamilies which underwent significant expansion in placental mammals. **b)** Family 1 subfamilies indicating potential orthologous pairs between the mammalian taxonomical classes.

labeled 1+7 are currently represented by more than 300 members, and families 2 and 5 have more than 80 each, while five other families have a current count of less than a dozen. A group of subfamilies within family 1 are significantly different from the other subfamilies and appear to have undergone significant expansions in the mammalian lineages. For convenience, and for compatibility with previous nomenclature, this subtree (Fig. 3a) is designated as family 7.

We produced a separate multiple alignment for each family, and then we constructed family-specific consensus sequences by using a simple majority rule at each position in the alignments. The neighbor-joining phylogenetic tree (not shown) of these consensus

Table 2. The marsupial and monotreme ORs discovered.

Marsupial (koala)				Monotreme (platypus)			
Trivial name	Family	Pseudo	Copies	Trivial name	Family	Pseudo	Copies
Koa2	1		3	Pla3	1	yes	2
Koa3	1			Pla4	6		
Koa5	7		2	Pla5	5		
Koa10	1			Pla7	1		
Koa13	1	yes	2	Pla11	7		
Koa20	1			Pla12	7	yes	
Koa23	1		4	Pla19	7		
Koa24	7	yes		Pla21	5	yes	2
Koa25	1		3	Pla23	7		2
Koa26	1		2	Pla24	5		
Koa27	1			Pla27	7	yes	4
Koa30	2		2	Pla28	7		3
Koa32	7		2	Pla29	7		3
Koa38	7			Pla30	1		
Koa41	1			Pla32	6		
Koa42	7	yes		Pla34	1		
Koa45	7			Pla36	7		
Koa46	5	yes		Pla38	7		2
Koa48	7						
Koa49	1		5				

The marsupial and monotreme ORs discovered: trivial name, family assignment, pseudogene status, and number of copies observed in the sample performed.

sequences indicates that some families are significantly related to each other (e.g., 2 and 10), but also that the branching order of most families is less conclusive. This supports the notion of a rapid radiation of ancestral OR genes at the time when OR families were established.

A small number of putative honeybee ORs have been cloned by nested-PCR amplification of honeybee cDNA with degenerate primers derived from mammalian sequences (Danty et al. 1994). These ORs belong in Class II and are more similar to mammalian ORs than any lamprey OR detected to date. We find it, therefore, extremely unlikely that these are bona fide invertebrate ORs, conserved since their divergence from vertebrates.

Monotreme and marsupial olfactory receptors. Fifty PCR subclones were sequenced for each of the platypus and koala DNA preparations. Of these, 29 and 36 reactions respectively yielded good quality sequence. Each identified OR gene was represented one to five times (Table 2), resulting in 18 and 20 unique OR sequences, respectively. All these sequences belonged to Class II, and for both species the majority was from families 1+7, while the rest were from families 2, 5, and 6. These results suggest a relatively small OR gene repertoire, less expanded than that of placental mammals. We have carried out an analysis based on the assumption that the number of appearances of a given receptor sequence is binomially distributed. This led to an estimated repertoire size of 50 for both species. Because of the large error introduced by the small sample size, PCR primer bias, and the low slope of the correlation curve beyond the observed maximum (not shown), this should be considered a minimal estimate. On the other hand, many of the observed monotreme and marsupial ORs are tightly related and appear to be later expansions (Fig. 3). This suggests that the OR repertoire of the early mammal was very small, not unlike that of fish (Ngai et al. 1993; Mombaerts 1999) and lampreys (Berghard and Dryer 1998). On the other hand, previous evidence (Glusman et al. 1996) suggests that the mammalian OR subfamilies include members that diverged from each other around the time of eutherian radiation. Thus, it may be the case that most of the OR expansion occurred over a relatively short evolutionary time period.

Nomenclature. A nomenclature system is proposed here, based on family and subfamily classification. It is consistent with currently

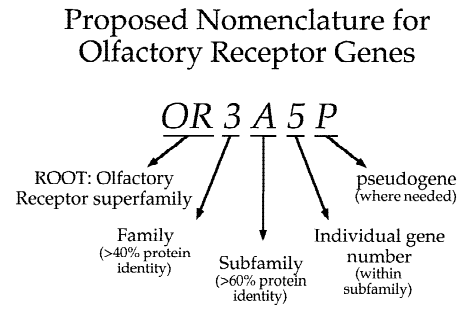

Fig. 4. Proposed nomenclature scheme for olfactory receptor genes. The gene name is to be preceded by a species tag.

accepted nomenclature schemes for other multigene families, e.g., the P450 superfamily (Nelson et al. 1996) and the UDP glucuronosyltransferase superfamily (Burchell et al. 1991) and with HUGO guidelines and recommendations (White et al. 1997, 1999).

The family and subfamily divisions and names were compared with those defined in a previous version of the OR nomenclature (Lancet and Ben-Arie 1993) and found to be generally compatible; family and subfamily designations were kept whenever possible. New subfamilies were introduced as needed. The letter 'O' was avoided for subfamily designation, to avoid confusion with the numeral '0'. Class II families were numbered starting at one (currently 1–14), while Class I families were numbered starting at 51 (currently 51–68).

The proposed OR nomenclature is based on the following five rules for building gene and product names (Fig. 4):

a) The italicized root symbol "OR" ("*Or*" for mouse), representing Olfactory or Odorant Receptors, is followed by an arabic number denoting the family, one or more letters denoting the subfamily, and an arabic number representing the individual gene within the subfamily. A hyphen should precede the final number in mouse genes.

b) "P" ("*ps*" in mouse) after the gene number denotes a pseudogene.

c) If a gene is the sole known member of a family or subfamily, the subfamily letter and gene number are included explicitly, to prevent unneeded redesignations when further sequences are made available.

d) The human nomenclature should be used for all species except mouse.

e) The name of the mRNA and protein products in all species (including mouse) should include all capital letters, without italics or hyphens.

Examples: "OR1E1" and "OR3A5P", for OR17-2 and ψOR17-25, respectively (Glusman et al. 2000).

Using these nomenclature guidelines, and on the basis of the classification results described above, we have assigned standardized names to the human OR genes detected in the databases. The proposed gene names have been submitted to the major databases (GDB/GenBank) and can be browsed via the HORDE web site (http://bioinfo.weizmann.ac.il/HORDE).

Nomenclature of orthologous genes. Orthologous genes in two or more species are those that we know, with a high degree of probability, correspond to the ancestral gene that existed before the evolutionary divergence of the two species. For example, a gene was described (Rouquier et al. 1998a) that is unique and shows conserved synteny in various primate genomes and whose evolutionary history could be clearly traced. On the other hand, any post-species-divergence events of gene duplication or conversion can make ortholog assignments difficult or equivocal. We have reported several such events in the primate evolutionary history of

the OR genes in the human Chr 17p13.3 cluster (Sharon et al. 1999). It is to be expected that orthology assignments for wider divergence times (e.g. human–rodent) should be more difficult, but some assignments could be made if OR clusters showing conservation of synteny are fully mapped or sequenced in more than one species (M. Lapidot et al., in preparation). It is expected that additional knowledge, confirming or denying orthology relations, may require gene redenominations. We, therefore, suggest that whenever orthology assignments are unclear, sequential numbering of new genes be followed, as in practice for the UDP glucuronosyltransferase gene superfamily (Burchell et al. 1991).

Closing remarks

Since the number of available OR sequences is growing quickly, there is an urgent need to adopt a standard naming method. We describe here the results of extensive data mining of available olfactory receptor gene sequences and their classification based on the analysis of their phylogenetic relationships. We hope the proposed nomenclature will be adopted widely and serve the community in providing a common ground for referring to olfactory receptor genes and gene products. We urge the researchers to consult with us for assigning names to newly described OR genes.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402

Asai H, Kasai H, Matsuda Y, Yamazaki N, Nagawa F et al. (1996) Genomic structure and transcription of a murine odorant receptor gene: differential initiation of transcription in the olfactory and testicular cells. Biochem Biophys Res Commun 221, 240–247

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

Berghard A, Dryer L (1998) A novel family of ancient vertebrate odorant receptors. J Neurobiol 37, 383–392

Buck L, Axel R (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65, 175–187

Buettner JA, Glusman G, Ben-Arie N, Ramos P, Lancet D et al. (1998) Organization and evolution of olfactory receptor genes on human chromosome 11. Genomics 53, 56–68

Bulger M, von Doorninck JH, Saitoh N, Telling A, Farrell C et al. (1999) Conservation of sequence and structure flanking the mouse and human beta-globin loci: the beta-globin genes are embedded within an array of odorant receptor genes. Proc Natl Acad Sci USA 96, 5129–5134

Burchell B, Nebert DW, Nelson DR, Bock KW, Iyanagi T et al. (1991) The UDP glucuronosyltransferase gene superfamily: suggested nomenclature based on evolutionary divergence. DNA Cell Biol 10, 487–494

Chess A, Simon I, Cedar H, Axel R (1994) Allelic inactivation regulates olfactory receptor gene expression. Cell 78, 823–834

Conticello SG, Pilpel Y, Glusman G, Fainzilber M (2000) Position-specific codon conservation in hypervariable gene families. Trends Genet 16, 57–59

Crowe ML, Perry BN, Connerton IF (1996) Olfactory receptor-encoding genes and pseudogenes are expressed in humans. Gene 169, 247–249

Danty E, Cornuet JM, Masson C (1994) Honeybees have putative olfactory receptor proteins similar to those of vertebrates. CR Acad Sci III 317, 1073–1079

Feingold EA, Forget BG (1989) The breakpoint of a large deletion causing hereditary persistence of fetal hemoglobin occurs within an erythroid DNA domain remote from the beta-globin gene cluster. Blood 74, 2178–2186

Feingold EA, Penny LA, Nienhuis AW, Forget BG (1999) An olfactory receptor gene is located in the extended human beta-globin gene cluster and is expressed in erythroid cells. Genomics 61, 15–23

Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166

Freitag J, Krieger J, Strotmann J, Breer H (1995) Two classes of olfactory receptors in *Xenopus laevis.* Neuron 15, 1383–1392

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

Freitag J, Beck A, Ludwig G, von Buchholtz L, Breer H (1999) On the origin of the olfactory receptor family: receptor genes of the jawless fish (*Lampetra fluviatilis*). Gene 226, 165–74

Glusman G, Clifton S, Roe R, and Lancet D (1996) Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity. Genomics 37, 147–60

Glusman G, Sosinsky A, Ben-Asher E, Avidan N, Sonkin D et al. (2000) Sequence, structure and evolution of a complete human olfactory receptor gene cluster. Genomics 63, 227–245

Griff IC, Reed RR (1995) The genetic basis for specific anosmia to isovaleric acid in the mouse. Cell 83, 407–414

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266, 383–402

Holm L, Sander C (1998) Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14, 423–429

Kosteas T, Palena A, Anagnou NP (1997) Molecular cloning of the breakpoints of the hereditary persistence of fetal hemoglobin type-6 (HPFH-6) deletion and sequence analysis of the novel juxtaposed region from the 3′ end of the beta-globin gene cluster. Hum Genet 100, 441–445

Lancet D (1986) Vertebrate olfactory reception. Annu Rev Neurosci 9, 329–355

Lancet D, Pace U (1987) The molecular basis of odor recognition. Trends Biochem Sci 12, 63–66

Lancet D, Ben-Arie N (1993) Olfactory receptors. Curr Biol 3, 668–674

Lancet D, Gross-Isseroff R, Margalit T, Seidmann E, and Ben-Arie N (1993a) Olfaction: from signal transduction and termination to human genome mapping. Chem Senses 18, 217–225

Lancet D, Sadovsky E, Seidemann E (1993b) Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. Proc Natl Acad Sci USA 90, 3715–3719

Lundin LG (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16, 1–19

Malnic B, Hirono J, Sato T, Buck LB (1999) Combinatorial receptor codes for odors. Cell 96, 713–723

Mombaerts P (1999) Molecular biology of odorant receptors in vertebrates. Annu Rev Neurosci 22, 487–509

Nef P, Hermans BI, Artieres PH, Beasley L, Dionne VE et al. (1992) Spatial pattern of receptor expression in the olfactory epithelium. Proc Natl Acad Sci USA 89, 8948–8952

Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R et al. (1996) P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics 6, 1–42

Ngai J, Dowling MM, Buck L, Axel R, Chess A (1993) The family of genes encoding odorant receptors in the channel catfish. Cell 72, 657–666

Page RDM (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12, 357–358

Pearson WR, Wood T, Zhang Z, Miller W. (1997) Comparison of DNA sequences with protein sequences. Genomics 46, 24–36

Qasba P, Reed RR (1998) Tissue and zonal-specific expression of an olfactory receptor transgene. J Neurosci 18, 227–236

Raming K, Konzelmann S, Breer H (1998) Identification of a novel G-protein coupled receptor expressed in distinct brain regions and a defined olfactory zone. Recept Channels 6, 141–151

Reed RR (1990) How does the nose know? Cell 60, 1–2

Rouquier S, Friedman C, Delettre C, van den Engh G, Blancher A et al. (1998a) A gene recently inactivated in human defines a new olfactory receptor family in mammals. Hum Mol Genet 7, 1337–1345

Rouquier S, Taviaux S, Trask BJ, Brand-Arpon V, van den Engh G et al. (1998b) Distribution of olfactory receptor genes in the human genome. Nat Genet 18, 243–250

Rouquier S, Blancher A, 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

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406–425

Sharon D, Glusman G, Pilpel Y, Khen M, Gruetzner F et al. (1999) Primate

evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. Genomics 61, 24–36

Sullivan SL, Adamson MC, Ressler KJ, Kozak CA, Buck LB (1996) The chromosomal distribution of mouse odorant receptor genes. Proc Natl Acad Sci USA 93, 884–888

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

Walensky LD, Ruat M, Bakin RE, Blackshaw S, Ronnett GV et al. (1998) Two novel odorant receptor families expressed in spermatids undergo 5′-splicing. J Biol Chem 273, 9378–9387

White JA, McAlpine PJ, Antonarakis S, Cann H, Eppig JT et al. (1997) Guidelines for human gene nomenclature (1997) HUGO Nomenclature Committee. Genomics 45, 468–471

White JA, Apweiler R, Blake JA, Eppig JT, Maltais LJ et al. (1999) Report on the Second International Nomenclature Workshop. Genomics 62, 320–323