Hum Genet (2001) 108 : 1–13
DOI 10.1007/s004390000436

ORIGINAL INVESTIGATION

**Tania Fuchs · Gustavo Glusman · Shirley Horn-Saban
Doron Lancet · Yitzhak Pilpel**

The human olfactory subgenome: from sequence to structure and evolution

Received: 10 July 2000 / Accepted: 27 October 2000 / Published online: 3 January 2001

Abstract Olfactory receptors (ORs) constitute the largest multigene family in multicellular organisms. Their evolutionary proliferation has been driven by the need to provide recognition capacity for millions of potential odorants with arbitrary chemical configurations. Human genome sequencing has provided a highly informative picture of the “olfactory subgenome”, the repertoire of OR genes. We describe here an analysis of 224 human OR genes, a much larger number than hitherto systematically analyzed. These are derived by literature survey, data mining at 14 genomic clusters, and by an OR-targeted experimental sequencing strategy. The presented set contains at least 53% pseudogenes and is minimally divided into 11 gene families. One of these (no. 7) has undergone a particularly extensive expansion in primates. The analysis of this collection leads to insight into the origin of OR genes, suggesting a graded expansion through mammalian evolution. It also allows us to delineate a structural map of the respective proteins. A sequence database and analysis package is provided (http://bioinformatics.weizmann.ac.il/HORDE), which will be useful for analyzing human OR sequences genome-wide.

Introduction

The olfactory pathway constitutes a remarkable molecular recognition apparatus, capable of detecting millions of volatile chemicals (Lancet 1986; Krieger and Breer 1999; Mombaerts 1999a, 1999b; Prasad and Reed 1999; Buck 2000). This is afforded by olfactory receptor (OR) proteins, a large repertoire of G-protein-coupled receptors expressed in specialized sensory neurons and present in most vertebrates. Homologs of the putative OR-coding genes, first discovered in rat (Buck and Axel 1991), are readily identified in other species by homology-based cloning or by sequence similarity searches (Parmentier et al. 1992; Selbie et al. 1992; Ngai et al. 1993; Ressler et al. 1993; Freitag et al. 1995; Barth et al. 1996; Issel-Tarver and Rine 1996; Nef et al. 1996; Sullivan et al. 1996; Trask et al. 1998a; Velten et al. 1998; Brand-Arpon et al. 1999; Rouquier et al. 2000).

The human OR gene repertoire (olfactory subgenome) is estimated to contain several hundred genes, but its accurate extent remains to be elucidated. OR-like sequences appear in dozens of genomic clusters on most chromosomes (Ben-Arie et al. 1994; Glusman et al. 1996, 2000b; Buettner et al. 1998; Rouquier et al. 1998; Trask et al. 1998a; Brand-Arpon et al. 1999), with a possible bias toward the chromosomal telomeres (Trask et al. 1998b; Fig. 1A). Such a cluster organization may reflect an evolutionary process of duplication of both individual genes and entire genomic segments (Glusman et al. 1996; Trask et al. 1998a; Brand-Arpon et al. 1999; Mombaerts 1999a) and could be crucial for the regulation of OR gene expression.

Extensive fluorescence in situ hybridization (FISH) analysis has revealed 44 human OR-containing genomic loci (Trask et al. 1998a; Fig. 1A). Previous publications have described several sequenced human OR clusters. These include a cluster of approximately 400 kb on chromosome 17p13 containing 17 genes (Ben-Arie et al. 1994; Glusman et al. 1996, 2000b), and a cluster of approximately 100 kb with four genes on chromosome 3p13 and which is duplicated on 3q13–21 (Brand-Arpon et al. 1999). Twenty five OR sequences have been reported in at least seven distinct regions of chromosome 11 (Buettner et al. 1998). In addition, compendia of partial OR-coding sequences have been reported by several laboratories (Parmentier et al. 1992; Rouquier et al. 1998).

The present paper describes a non-redundant collection of 224 OR-coding sequences stemming from cloning and

T. Fuchs and G. Glusman contributed equally to this work

T. Fuchs · G. Glusman · S. Horn-Saban · D. Lancet (✉) · Y. Pilpel
Department of Molecular Genetics
and the Crown Human Genome Center,
The Weizmann Institute of Science,
Rehovot 76100, Israel
e-mail: doron.lancet@weizmann.ac.il,
Tel.: +972-8-9343683 or +972-8-9344121, Fax: +972-8-9344112

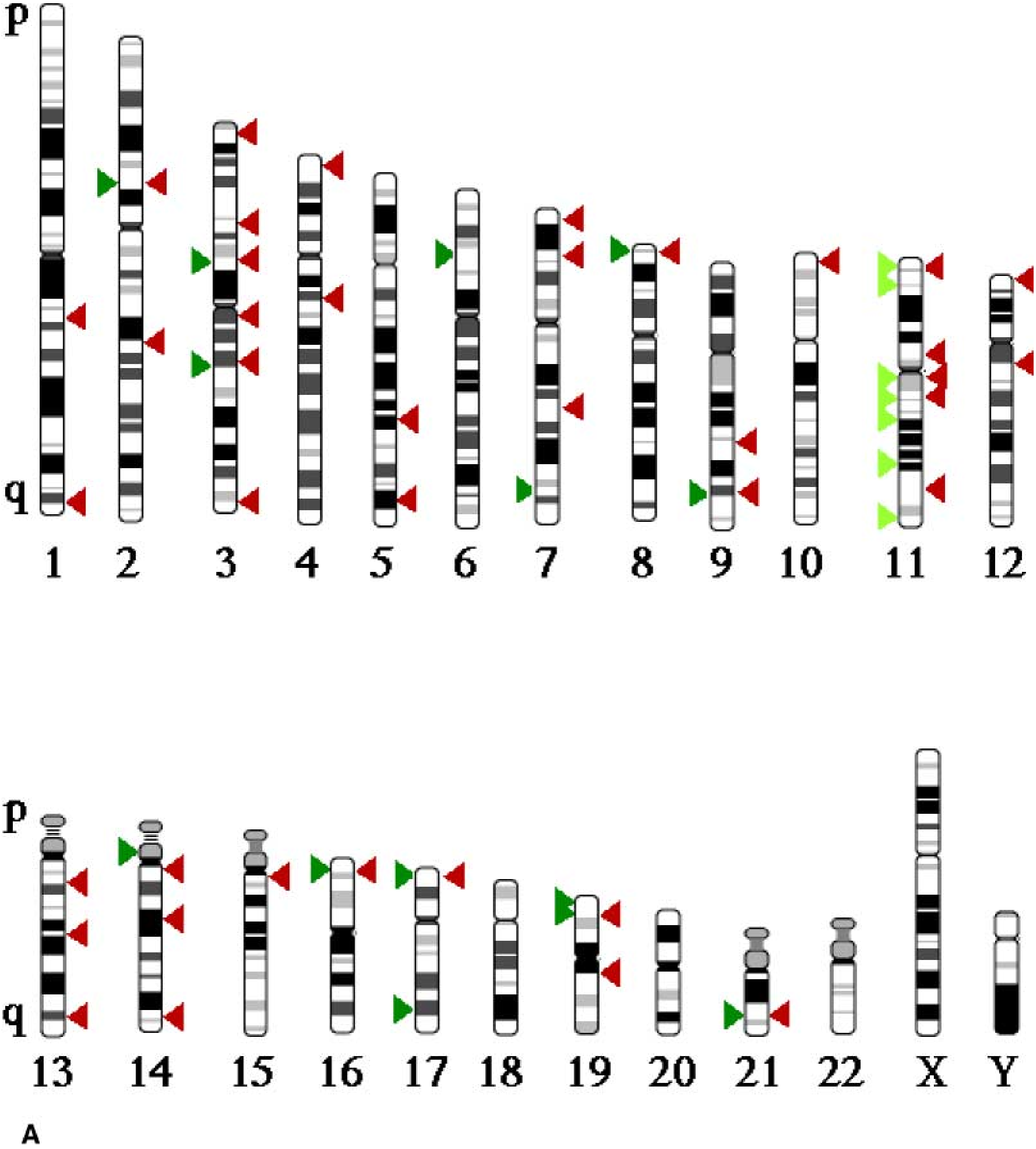


Fig. 1 **A** The chromosomal locations of OR gene clusters in the human genome. *Red triangles* Localization by fluorescence-based in situ hybridization with mixtures of OR-containing PCR products (Rouquier et al. 1998) or OR-containing genomic clones (Trask et al. 1998a), *light green triangles* identified OR clusters that have not undergone genomic sequencing, *dark green triangles* genomically sequenced OR clusters. The chromosomal localization of these clones is according to the GenBank annotation, except for the clusters at 2p13, 7q33, 9q33, 14p12, and 17q23 for which locations are based on electronic PCR (e-PCR; Schuler 1997) in conjunction with Unified Database (UDB) coordinates (Chalifa-Caspi et al. 1997; http://bioinformatics.weizmann.ac.il/udb/). Independent data from single chromosome PCR amplification (Rouquier et al. 1998) are also consistent with the presence of OR genes on chromosomes 1–21 and X, but not 22 or Y. **B** Human OR clusters for which genomic sequences are available. *Boxes above* and *under* the *line* indicate the absolute orientation of ORs to each other. Cluster sizes and intergenic distances are shown to scale. *Red boxes* OR pseudogenes, *green boxes* OR genes. Family and subfamily affiliation are indicated for each gene according to the nomenclature system of Glusman et al. (2000a). The cluster at 7q33–35 is composed of three separate sequence contigs, two of which are localized by e-PCR/UDB analysis to the same cytogenetic region as the one localized to this region by GenBank annotation

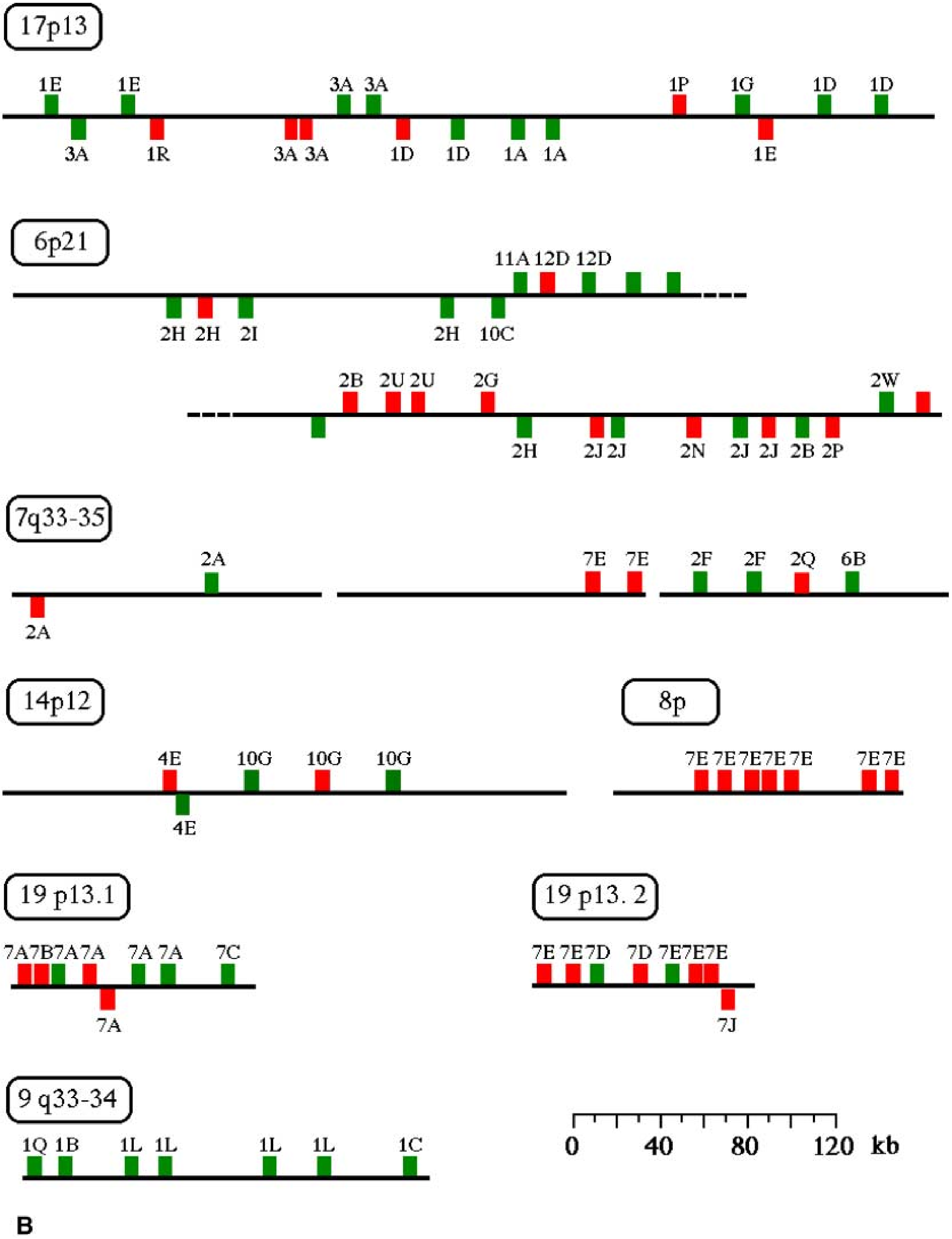

17p13
1E 1E 3A 3A 1P 1G 1D 1D
3A 1R 3A 3A 1D 1D 1A 1A 1E
6p21
11A 12D 12D
2H 2H 2I 2H 10C
2B 2U 2U 2G 2W
2H 2J 2J 2N 2J 2J 2B 2P
7q33-35
2A 7E 7E 2F 2F 2Q 6B
2A
14p12
4E 10G 10G 10G
4E
8p
7E 7E 7E 7E 7E 7E 7E
19 p13.1
7A 7B 7A 7A 7A 7A 7C
7A
19 p13. 2
7E 7E 7D 7D 7E 7E 7E
7J
9 q33-34
1Q 1B 1L 1L 1L 1L 1C
0 40 80 120 kb
B

Fig. 1 B

sequencing experiments and from data mining. This coverage is instrumental for extracting new information concerning the olfactory gene superfamily.

Materials and methods

Generation of olfactory sequence tags

Polymerase chain reaction (PCR) at low stringency (annealing temperature –55°C) was performed on the genomic DNA of one or more individuals by means of OR-specific degenerate primers 5B: CCCATGTAYTTBTTYCTCDSYAAYYTRTC (corresponding to TM2) and 3B: ATGNTGAAYCCNTTCATNTAYWGYCT (corresponding to TM7; Ben-Arie et al. 1994) modified for subsequent subcloning into the pAMP vector. PCR products were subcloned into the pAMP vector by using the Clone Amp System as described by the manufacturer (Gibco Brl). PCR with vector-specific primers was performed on subclones. PCR products were subjected to an automatic sequencing procedure (ABI 377). Forward and reverse sequences were assembled for each clone and visually corrected by using the computer program Sequencher version 3.0 (Gene Codes corporation). Sequences corresponding to the consensus primers 3B and 5B were removed at both ends of each sequence. Sequence identification was performed by BLAST (basic local alignment search tool) against the human olfactory receptor specific database (http://bioinformatics.weizmann.ac.il/HORDE).

Sequence analysis

A series of BLAST searches against the GenBank (~20% of the human genome) was performed by using various OR gene sequences as queries. The sequences of the large-scale clones were retrieved and subjected to BLAST against human OR-specific database to retrieve regions corresponding to OR sequences.

An electronic PCR (e-PCR) program for finding sequence tagged sites (STSs) in DNA sequences (Schuler 1997) was applied to determine the genomic localization of clones. According to the STSs found, it is possible to estimate the chromosomal position of the clone by using the Unified Database (UDB) maps (Chalifa-Caspi et al. 1997).

The obtained OR sequences were translated conceptually by using FASTY (Pearson et al. 1997) against a "core" of properly translated OR sequences, i.e., sequences that we trust with a high degree of confidence are translated in the correct reading frame. This translation procedure ensures that pseudogenes bearing frameshift mutations are corrected in a fashion that allows their proper alignment with intact OR genes.

Multiple alignment of the deduced amino acid sequences of ORs was performed by a fully automated application of the ClustalW program (Thompson et al. 1994; Higgins et al. 1996) with default parameters.

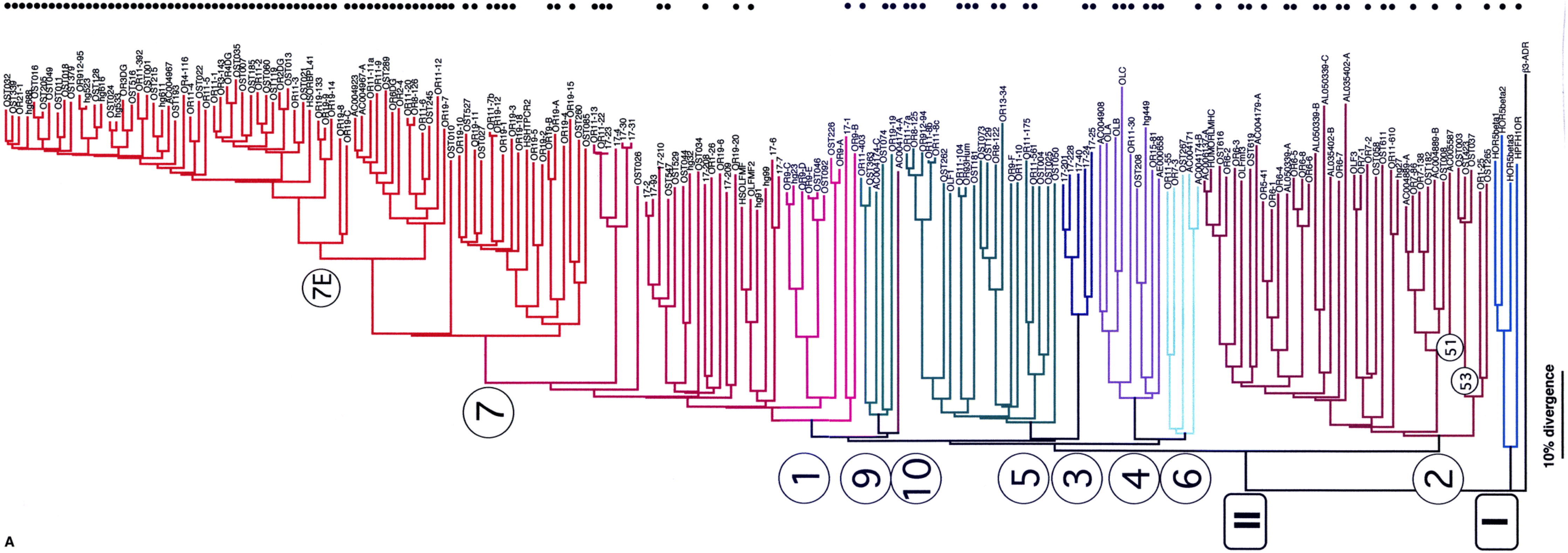
Results

Previous publications have included more than 100 complete or partial human OR-coding region sequences (Penny et al. 1989; Ben-Arie et al. 1994; Crowe et al. 1996; Fan et al. 1996; Glusman et al. 1996; Issel-Tarver et al. 1997; Kosteas et al. 1997; Vanderhaeghen et al. 1997; Buettner et al. 1998; Rouquier et al. 1998). We have initiated an effort targeted at obtaining information about additional OR genes in the human genome. An interim search in high throughput genomic sequence data identified 59 full-length OR sequences in 14 clusters (Fig. 1A, B). Some of them also appear in the Olfactory Receptor Data Base (Skoufos et al. 2000). The largest cluster so far is on chromosome 6p21 in the vicinity of the major histocompatibility complex locus (Fan et al. 1996; Ziegler 1997) and contains 27 OR genes. The second largest cluster on chromosome 17p13 contains 17 OR sequences within a 412-kb region (Ben-Arie et al. 1994; Glusman et al. 1996, 2000b). Two additional clusters on chromosome 19 are being characterized in more detail (S. Horn-Saban, L. Ashworth et al., unpublished).

In parallel, we have initiated an effort to sequence new OR-coding regions from the entire human genome. This was carried out by PCR amplification on genomic DNA by using OR-specific degenerate primers corresponding to the conserved regions in transmembrane helices 2 and 7 (TM2 and TM7) of the OR protein (Ben-Arie et al. 1994). These are termed Olfactory Sequence Tags (OSTs), as they can be used to identify the complete genomic sequences derived from automated data mining. Because of the obvious danger of sequencing artefacts, a special effort was made to avoid chimerism and other sequencing errors (see Materials and methods). An analysis of 390 such clones revealed 64 novel OR sequences. Of these, 39 were found to match with more than 99% identity with genomic sequences (G. Glusman, unpublished). Of the remaining sequences, none revealed the existence of a chimerism.

Based on these two resources and on previous publications, a non-redundant set of 224 human olfactory receptor coding sequences (cutoff at 99% amino acid identity) is presented here. This collection only includes sequences minimally extending between TM2 and TM7 (Fig. 1B). Additional information referring to this collection is available at http://bioinformatics.weizmann.ac.il/HORDE (under "publications"). The site also contains various tools for OR gene and protein analysis. In 129 cases, the full-length coding region and the accurate genomic localization are reported. About 50 previously published shorter sequences (Vanderhaeghen et al. 1997) were not included in the present analyses. The number of genes represented by these 224 sequences may be larger, because of the existence of more than one copy of the same sequence within large-scale genomic duplications (Trask et al. 1998a). Based on very rough estimates of the size of the human olfactory subgenome (500–1000 genes; Buck and Axel 1991; Lancet et al. 1993b), the current OR collection may constitute between one quarter and one half of the entire genomic gamut.

Within the entire collection of 224 OR sequences, 119 (53%) show frame disruptions (frameshifts and/or in-frame stop codons) suggesting that they are pseudogenes (Fig. 2A,B). Among the presently reported full-length sequences, 60 are pseudogenes, of which 47 have deleterious changes in their TM2–TM7 segment. Based on this knowledge, it is possible to compute an extrapolated number of potential pseudogenes among the ORs for which the full-length sequence is not presently reported. Therefore, the 59 OR partial sequences identified as pseudogenes may represent an actual number of 75. Thus, the total extrapolated pseudogene count in the reported collection is 135, and the overall extrapolated percentage is 60%.


7E
7
1
9
10
5
3
4
6
2
51
53
II
I
β3-ADR
10% divergence
A

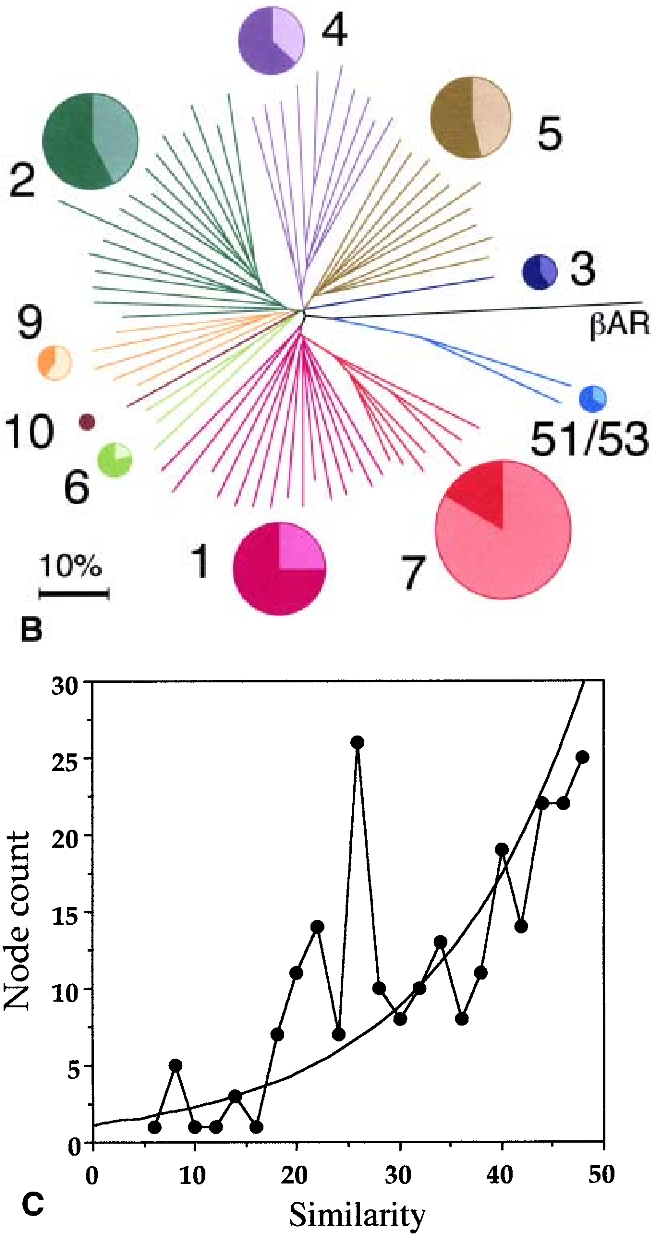


Fig. 2 **A** A similarity dendrogram of all 224 human OR sequences presented. OR gene families are denoted by *circled numbers* and are individually colored. Pseudogenes are marked with a *dot* near their trivial name. The *10% divergence bar* indicates the corrected protein sequence difference summed over the distance separating a given gene pair. The analysis is based on a conceptual translation of DNA sequences for OR coding regions sequenced at least between the 2nd to the 7th transmembrane helices (TM2–TM7), and that bear less than 99% pairwise identity to each other. The identity scores are computed only for TM2–TM7 region, shared by all sequences. The sequences were multiply aligned with ClustalW (Higgins et al. 1996), and the dendrogram was obtained with the TreeView program (Page 1996). The family and subfamily classifications were determined by a computerized optimization procedure (Glusman et al. 2000a). **B** A similarity dendrogram of consensus sequences of each of the 73 OR subfamilies, with family *identification numbers*. The area of the circles is proportional to the number of genes in each family. *Dark* and *light colors* respectively indicate the percentage of genes and pseudogenes in each OR family. The human beta-3 adrenergic receptor (ADRB3) served as an outgroup. All families are class II except for families 51/53, which are class I (fish-like). A *10% divergence bar* is shown. The 224 OR-like protein sequences were obtained by conceptual translation of the original DNA sequences. The OR collection included only those for which the sequence was available at least from the 2nd to the 7th transmembrane helices and only sequences that bore less than 99% identity to each other. The analyses were performed for segments between the 2nd to the 7th transmembrane helices shared by all sequences. The consensus sequences were multiply aligned with ClustalW (Higgins et al. 1996), and the dendrogram was obtained with the TreeView program (Page 1996). The family and subfamily classifications were determined by a computerized optimization procedure (Glusman et al. 2000a). **C** An exponential increase of the node count in the OR dendrogram. For each of the 223 nodes (bifurcation points) in the dendrogram containing 224 sequences, a similarity parameter was computed as S=50–0.5*(100–I)=0.5 I, where I is the average pairwise similarity among all sequences stemming from that node. The experimental points indicate node counts (C), i.e., the number of nodes that have S values within a specified range. The smooth line is the curve-fitted exponent C=1.18*$10^{0.029S}$. An increment of S=10 (equal to an I increment of 20%) corresponds to an increase of C by a factor of $10^{0.29}\cong 2$. We have previously shown (Glusman et al. 1996) that two OR sequences with mutual similarity of 20% diverged 80–100 million years ago. Thus, it is inferred that the doubling time of the olfactory repertoire is roughly within this time range

The 224 OR coding sequences were subjected to phylogenetic analysis based on overall sequence similarity (Fig. 2A). By using standard definitions (Dayhoff 1976), the sequences encompass 11 families and 73 subfamilies. A nomenclature system based on such a classification has been previously proposed (Lancet and Ben-Arie 1993; Glusman et al. 2000a). This analysis is traditionally performed on the most informative part of the sequence, one that is also available for a larger fraction of the repertoire in multiple species, and has led to the current nomenclature for OR genes (Glusman et al. 2000a). Such previous analyses have indicated that there is no major loss of classification information when such partial sequences are employed.

The complete dendrogram was subjected to a duplication time analysis (Fig. 2C). For each of the nodes (bifurcation points) in the neighbor-joining tree, a similarity parameter was computed based on the average pairwise similarity among all sequences stemming from the node. Under the assumption of an average constant rate of mutation along all tree branches, this parameter can be assumed to relate linearly to divergence time. The number of bifurcation points (nodes) in the tree have been found to increas roughly exponentially with respect to a function of the increasing percent similarity. A parameter fit suggests that, on average, the repertoire size may have doubled approximately every 80 million years. A back-extrapolated OR count of roughly 25 is computed for the arrival of air-breathing vertebrates approximately 350 million years ago.

Analyses of the protein sequence of all 224 ORs reveal a highly conserved protein sequence (Fig. 3). Seventy two residues (amounting to about a quarter of the entire protein length) are above the 80% consensus line. Such conservation patterns are better highlighted and more statistically evident now that so many OR sequences are avail-

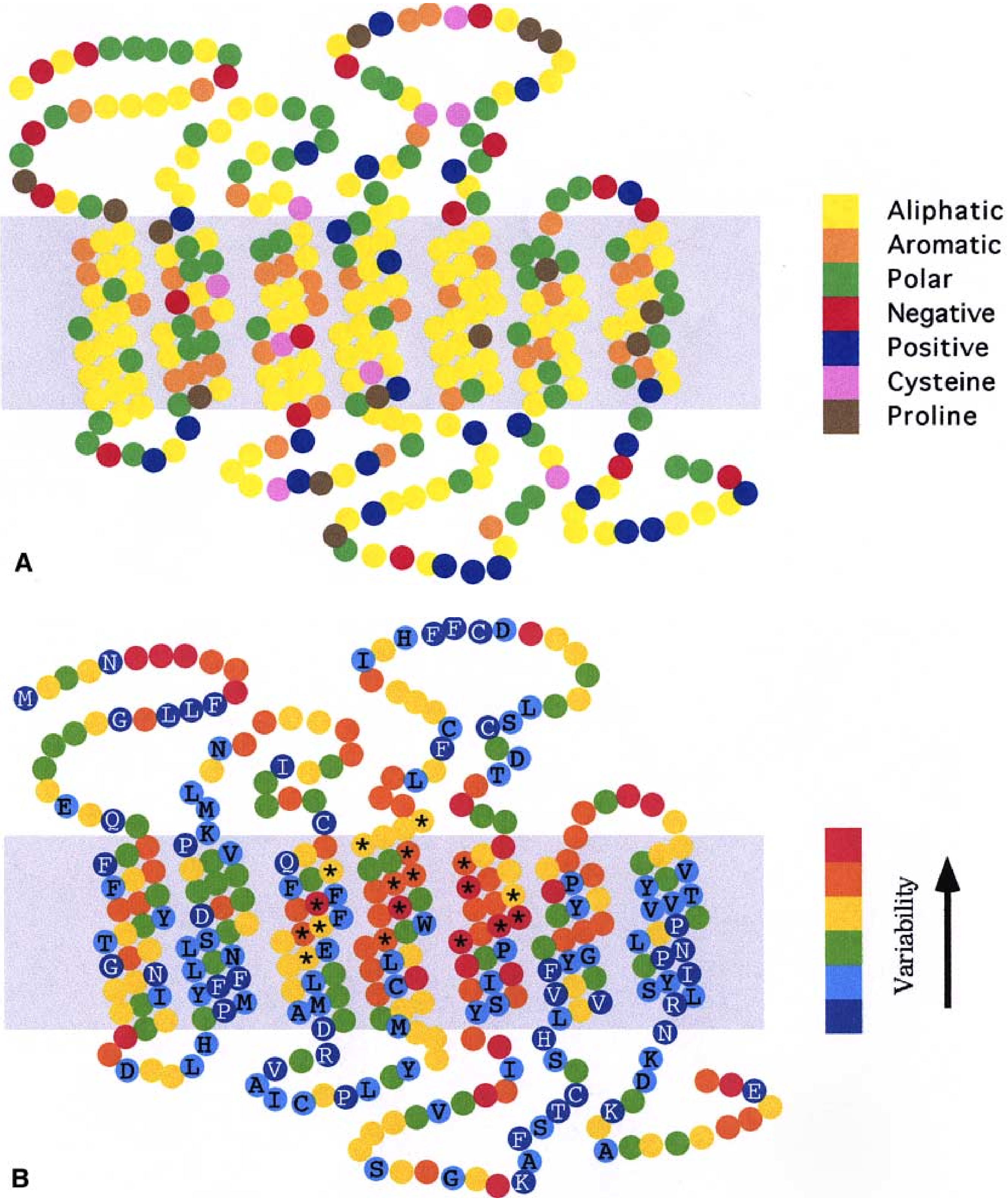


Fig. 4A, B Schematic diagram of the OR protein structure, with the predicted location of the membrane as a *shaded box*. **A** Amino acid residue positions of the OR consensus sequence are grouped and color-coded according to their types as in Fig. 3. **B** Amino acid variability index (Pilpel and Lancet 1999) calculated for each *column* in the multiple alignment of all 224 human ORs, grouped and color-coded with 6 degrees of diversity. Residues in the two most conserved levels are indicated by *single letter code*. In addition, the 17 hypervariable complementarity determining region (CDR) residues of the OR protein (Pilpel and Lancet 1999) are marked by *asterisks*

9 10

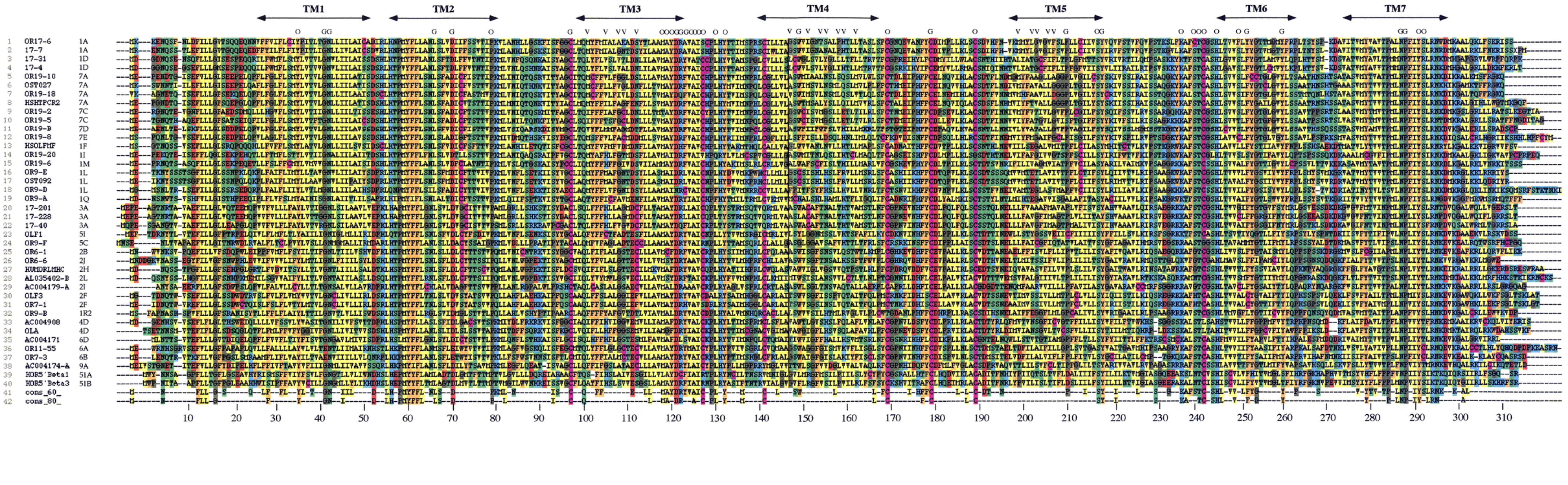


Fig. 3 Multiple alignment of human OR proteins. The 40 sequences shown are of OR coding regions that are fully sequenced and include no pseudogenes. They include representatives of all known human OR gene families. The rows *cons_60* and *cons_80* are the consensi of all 69 fully sequenced open reading frames, calculated by 60% and 80% plurality, respectively. The seven putative transmembrane helices are shown on *top*. The following positions are marked *above* the sequences: *V* residue of a putative odorant-binding site, *G* conserved position among all G-protein-coupled receptors, *O* highly conserved positions unique to ORs. *Colors* indicate the amino acid types: *red* acidic, *blue* basic, *green* uncharged polar, *yellow* aliphatic hydrophobic, *orange* aromatic, *brown* proline/glycine, *purple* cysteine

able from one vertebrate species. Among the most notable positions are seven conserved cysteines, some of which could play a role in maintaining the structural integrity of this 7TM protein. Whereas two of these cysteines (at positions 97 and 179) are common to all G-protein coupled receptors (GPCRs) and are believed to form a disulfide-link between extracellular loops 1 and 2 (Baldwin 1994), the other five are unique to ORs. Two (Cys 170 and Cys 189) may form another cysteine bridge, conceivably important for the conformation of extracellular loop 2, a domain previously proposed potentially playing a role in olfactory axonal guidance (Singer et al. 1995). At least two of the other three are intracellular and might therefore be maintained in a reduced unbridging state. Some of the conserved cysteines could underlie the known sensitivity of the olfactory apparatus to thiols (Theimer et al. 1977; Lancet 1986).

The length of the OR protein is highly stereotyped (313±8 residues) suggesting functional constraints that may prohibit larger insertions and deletions. Most of the length deviations are in the amino- and carboxy-terminals, and rare insertions (1–16 amino acids) can be found throughout the protein length. The latter practically always occur in the loops that connect the transmembrane helices. The seven transmembrane helices occupy more than half of the total length of the short OR polypeptide (Fig. 3). Nevertheless, the entire protein is less lipophilic than expected; its longest uninterrupted run of hydrophobic residues is seven amino acids long. Some of the presumed transmembrane helices are atypical, being interrupted by polar and charged residues, and by helix-kinking prolines. No obvious signal sequence is seen, but a prominent, highly conserved NXS/T consensus for N-linked glycosylation is found near to the amino terminal.

In general, the intracellular half of the protein is more positively charged and more conserved than the extracellular portion (Fig. 4A, B). The intracellular loops, which probably form the G-protein interface of the OR protein, contain some of the most immutable consensus segments, rich in positively charged amino acids and in serine and threonine residues, which are potential sites for phosphorylation (Pilpel and Lancet 1999; Fig. 4B). In contrast, the extracellular loops, in particular the first and the third, and the N- and C-terminals, show a relatively high inter-OR variability.

Discussion

Two of the human OR families (51 and 53), presently encompassing six genes, show a significant resemblance to OR genes in fish (Lancet and Ben-Arie 1993; Freitag et al. 1995). These ORs have been termed class I or “fish-like”, in contrast to the rest of the OR families that are known as class II (Freitag et al. 1995; Fig. 2A, B). The human class I receptors may therefore be relics of an ancient group of OR genes that survived alongside with the much expanded set of air-breathing-type class II genes.

It appears that the olfactory subgenome has expanded steadily throughout vertebrate evolution. This is consistent with results of our data mining, indicating that most of the OR families present in humans are also found in other vertebrates as far back as amphibians (Freitag et al. 1995; Glusman et al. 1996, 2000a). This finding is also in line with the notion that air-breathing vertebrates have evolved an OR repertoire distinct from that of fish (Lancet and Ben-Arie 1993; Freitag et al. 1995).

Interestingly, the duplication time analysis depicted in Fig. 2C also indicates a local deviation from the overall exponential trend, forming a peak of gene duplication at an estimated time period around 200 million years ago, corresponding to the period of reptile dominance. The significance of such inferred “OR radiation” will have to be elucidated further when the complete human olfactory subgenome is deciphered.

The general trend of exponential expansion appears to be true only on average. A closer examination of the entire OR dendrogram (Fig. 2A) shows considerable asymmetries: some families (e.g., families 2 and 5) appear to have entered a quiescent period long ago, as indicated by their large subfamily count and small number of genes per subfamily (average of 3 ORs). On the other hand, family 7 has very large subfamilies, with an average of 12 ORs per subfamily and one subfamily (7E) having nearly 50 members (Fig. 2A, B). This suggests a family- and subfamily-specific duplication activity, particularly during the more recent periods of mammalian evolution.

The reported pseudogene fraction (53%–60%) is somewhat lower than the 70% previously reported (Rouquier et al. 1998), a difference potentially stemming from the larger statistics in the present data and differences in the methodology for redundancy elimination. The average prevalence of pseudogenes in families 1–6, 9, and 10 is 37%, a high value, but not completely out of line with that seen in other multigene families (McCormack et al. 1993). In contrast, family 7 has a very high pseudogene incidence with only 15% of the ORs seemingly intact. Some of these inactive genes (mainly subfamily 7E) may have formed directly by the duplication of an ancestral pseudogene, as they share a common 4-bp deletion. Yet others appear to have been individually inactivated, since the rest of family 7 still has a high pseudogene count (73%). This may be an extreme example of the general OR gene loss suggested to have occurred during primate evolution (Sharon et al. 1999; Rouquier et al. 2000).

Two basic models have been proposed for the expansion of the OR repertoire (Lancet et al. 1993a; Glusman et al. 1996; Trask et al. 1998a; Brand-Arpon et al. 1999): one involves duplication of individual genes, and the other, the replication of entire OR clusters. Some of the OR clusters are heterogeneous and contain representatives of several families and subfamilies (Fig. 1B). Thus, a subfamily-specific expansion may be indicative of a mechanism (e.g., recombination hotspots) that resides near individual genes. This, however, does not preclude that, in some cases, entire clusters would be duplicated, as shown experimentally (Trask et al. 1998a; Brand-Arpon et al.

1999). This is also corroborated by examining a pairwise distance distribution within and among clusters (not shown), which shows a bimodality for the intra-cluster pairs, suggesting a small but significant bias for within-cluster duplication.

The lowest pairwise similarities observed in the analyzed OR collection are 20%–25% identity (Fig. 2A, B). These low values are seen for comparisons across class and remote family boundaries. Such low scores are still sufficient to define unequivocally any newly sequenced coding regions such as OR genes and to distinguish OR genes from other superfamilies of chemosensory receptor genes.

Of special interest is the potentially functional variability seen in some of the transmembrane helices (Fig. 4B). The binding of odorous ligands to OR proteins is likely to occur in hypervariable regions, as is the case for antigen binding to immunoglobulins. On the other hand, for many GPCRs, a pocket within the transmembrane helix barrel, mainly delineated by TMs 2–6, has been shown to be involved in ligand recognition (Baldwin 1994; Herzyk and Hubbard 1995). Combining these criteria based on an OR collection from several species (Pilpel and Lancet 1999), we have recently identified 17 hypervariable residues in the extracellular two-thirds of TM segments 3, 4, and 5 as candidate complementarity determining regions (CDRs) in the OR proteins. The present analysis corroborates the proposed model, since the CDR positions also show high variability in the extensive human-only OR arsenal (Fig. 4B).

In conclusion, the olfactory system constitutes a most interesting model for understanding molecular recognition, signal transduction, neuronal information processing, and multigene family evolution. The present analysis of a considerable portion of the OR gene superfamily provides insights into structure and evolution. It is expected that, in the future, a fully deciphered OR repertoire should enhance and expand these results.

References

Baldwin J (1994) Structure and function of receptors coupled to G proteins. Curr Opin Cell Biol 6:180–190

Barth AL, Justice NJ, Ngai J (1996) Asynchronous onset of odorant receptor expression in the developing zebrafish olfactory system. Neuron 16:23–34

Ben-Arie N, Lancet D, Taylor C, Khen M, Walker N, Ledbetter DH, 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

Brand-Arpon V, Rouquier S, Massa H, Jong PJ de, Ferraz C, Ioannou PA, Demaille JG, Trask BJ, Giorgi D (1999) A genomic region encompassing a cluster of olfactory receptor genes and a myosin light chain kinase (MYLK) gene is duplicated on human chromosome regions 3q13-q21 and 3p13. Genomics 56: 98–110

Buck LB (2000) The molecular architecture of odor and pheomone sensing in mammals. Cell 100:611–618

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, Evans GA (1998) Organization and evolution of olfactory receptor genes on human chromosome 11. Genomics 53:56–68

Chalifa-Caspi V, Rebhan M, Prilusky J, Lancet D (1997) The Unified Database (UDB): a novel genome integration concept. Genome Digest 4:15

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

Dayhoff M (1976) The origin and evolution of protein superfamilies. Fed Proc 35:2132–2138

Fan W, Cai W, Parimoo S, Schwarz DC, Lennon GG, Weissman SM (1996) Identification of seven new human MHC class I region genes around the HLA-F locus [published erratum appears in Immunogenetics (1997) 46:169]. Immunogenetics 44:97–103

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

Glusman G, Clifton S, Roe B, Lancet D (1996) Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity. Genomics 37:147–160

Glusman G, Bahar A, Sharon D, Pilpel Y, White J, Lancet D (2000a) The olfactory receptor gene superfamily: data mining, classification and nomenclature. Mamm Genome 11:1016–1023

Glusman G, Sosinsky A, Ben-Asher E, Avidan N, Sonkin D, Bahar A, Rosenthal A, Clifton S, Roe B, Ferraz C, Demaille J, Lancet D (2000b) Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. Genomics 63:227–245

Herzyk P, Hubbard RE (1995) Automated method for modeling seven helix transmembranal receptors from experimental data. Biophysical J 69:2419–2442

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

Issel-Tarver L, Rine J (1996) Organization and expression of canine olfactory receptor genes. Proc Natl Acad Sci USA 93: 10897–10902

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

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

Krieger J, Breer H (1999) Olfactory reception in invertebrates. Science 286:720–723

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

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

Lancet D, Ben-Arie N, Cohen S, Gat U, Gross-Isseroff R, Horn-Saban S, Khen M, Lehrach H, Natochin M, North M, Seidemann E, Walker N (1993a) Olfactory receptors: transduction, diversity, human psychophysics and genome analysis. In: Goody J (ed) The molecular basis of smell and taste transduction. CIBA Foundation, London, pp 131–146

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

McCormack WT, Hurley EA, Thompson CB (1993) Germ line maintenance of the pseudogene donor pool for somatic immunoglobulin gene conversion in chickens. Mol Cell Biol 13:821–830

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

Mombaerts P (1999b) Odorant receptor genes in humans. Curr Opin Genet Dev 9:315–320

Nef S, Allaman I, Fiumelli H, De Castro E, Nef P (1996) Olfaction in birds: differential embryonic expression of nine putative odorant receptor genes in the avian olfactory system. Mech Dev 55:65–77

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. Computer Appl Biosci 12:357–358

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

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

Penny LA, Forget BG (1989) A conserved gene 3’ to the HPFH-1 deletion breakpoint may have an effect on fetal globin gene expression in HPFH 1. Prog Clin Biol Res 316B:133–141

Pilpel Y, Lancet D (1999) The variable and conserved interfaces of modeled olfactory receptor proteins. Protein Sci 8:969–977

Prasad BC, Reed RR (1999) Chemosensation: molecular mechanisms in worms and mammals. Trends Genet 15:150–153

Ressler KJ, Sullivan SL, Buck LB (1993) A zonal organization of odorant receptor gene expression in the olfactory epithelium. Cell 73:597–609

Rouquier S, Taviaux S, Trask BJ, Brand-Arpon V, Engh G van den, Demaille J, Giorgi D (1998) 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

Schuler GD (1997) Sequence mapping by electronic PCR. Genome Res 7:541–550

Selbie LA, Townsend NA, Iismaa TP, Shine J (1992) Novel G protein coupled receptors: a gene family of putative human olfactory receptor sequences. Mol Brain Res 13:159–163

Sharon D, Glusman G, Pilpel Y, Khen M, Gruetzner F, Haaf T, Lancet D (1999) Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. Genomics 61:24–36

Singer MS, Shepherd GM, Greer CA (1995) Olfactory receptors guide axons. Nature 377:19–20

Skoufos E, Marenco L, Nadkarni P, Miller P, Shepherd G (2000) Olfactory receptor database: a sensory chemoreceptor resource. Nucleic Acids Res 28:341–343

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

Theimer E, Yoshida T, Klaiber E (1977) Olfaction and molecular shape. Chirality as a requisite for odor. J Agric Food Chem 25:1168–1177

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

Trask B, Massa H, Brand-Arpon V, Chan K, Friedman C, Nguyen O, Eichler E, Engh G van den, Rouquier S, Shizuya H, Giorgi D (1998a) Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. Hum Mol Genet 7:2007–2020

Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, Collins C, Giorgi D, Iadonato S, Johnson F, Kuo WL, Massa H, Morrish T, Naylor S, Nguyen OT, Rouquier S, Smith T, Wong DJ, Youngblom J, Engh G van den (1998b) 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

Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M (1997) Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species. Genomics 39:239–246

Velten F, Rogel-Gaillard C, Renard C, Pontarotti P, Tazi-Ahnini R, Vaiman M, Chardon P (1998) A first map of the porcine major histocompatibility complex class I region. Tissue Antigens 51:183–194

Ziegler A (1997) Biology of chromosome 6. DNA Seq 8:189–201