Research

The human olfactory receptor repertoire

Sergey Zozulya, Fernando Echeverri and Trieu Nguyen

Address: Senomyx Inc., 11099 North Torrey Pines Road, La Jolla, CA 92037, USA.

Correspondence: Sergey Zozulya. E-mail: sergey.zozulya@senomyx.com

Published: 1 June 2001

Genome ***Biology*** 2001, **2(6)**:research0018.1–0018.12

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/6/research/0018

Received: 8 March 2001
Revised: 12 April 2001
Accepted: 18 April 2001

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: The mammalian olfactory apparatus is able to recognize and distinguish thousands of structurally diverse volatile chemicals. This chemosensory function is mediated by a very large family of seven-transmembrane olfactory (odorant) receptors encoded by approximately 1,000 genes, the majority of which are believed to be pseudogenes in humans.

Results: The strategy of our sequence database mining for full-length, functional candidate odorant receptor genes was based on the high overall sequence similarity and presence of a number of conserved sequence motifs in all known mammalian odorant receptors as well as the absence of introns in their coding sequences. We report here the identification and physical cloning of 347 putative human full-length odorant receptor genes. Comparative sequence analysis of the predicted gene products allowed us to identify and define a number of consensus sequence motifs and structural features of this vast family of receptors. A new nomenclature for human odorant receptors based on their chromosomal localization and phylogenetic analysis is proposed. We believe that these sequences represent the essentially complete repertoire of functional human odorant receptors.

Conclusions: The identification and cloning of all functional human odorant receptor genes is an important initial step in understanding receptor-ligand specificity and combinatorial encoding of odorant stimuli in human olfaction.

Background

Olfaction is a major neurosensory function by which mammals investigate the external chemical environment. The initial step in odor identification is interaction of an odorant molecule with olfactory (odorant) receptors (ORs) expressed at the surface of cilia of chemosensory olfactory neurons in the olfactory epithelium. Seven-transmembrane ORs, first identified in 1991 [1], are the largest vertebrate gene family, comprising as many as 1,000 genes (reviewed in [2-6]). Mammalian ORs are classical G-protein-coupled receptors belonging to Class I or A, which also includes opsins and catecholamine receptors [7]. Each olfactory neuron appears to express a single type of OR [8-10] implying a sophisticated mechanism of OR gene choice. Another intriguing feature of olfaction is combinatorial recognition of odorants. Each receptor recognizes multiple odorants, and each odorant binds to multiple receptors to generate specific activation patterns for each of a vast number of distinct smells [10].

The genes encoding ORs are devoid of introns within their coding regions [1,11]. Mammalian OR genes are typically organized in clusters of ten or more members and located on many chromosomes [12-14]. The repertoire of human OR (hOR) genes contains a large fraction of pseudogenes, suggesting that olfaction became less important in the course of primate evolution. Recent studies indicate that some 70% of all hOR genes may be pseudogenes, compared with fewer than 5% in rodents or lower primates [15,16].

Analyses of incomplete compilations of hORs, in particular approximately 150 full-length receptor genes [17,18], have recently been published. A larger annotated set of hOR genes is available as an online database [19]. The very recent milestone publication of the first draft of the human genome sequence by two groups [20,21] opens up the possibility of detailed and complete identification, mapping and analysis of OR genes and their products in the near future. One of these groups reported that the human genome contains 906 OR genes, of which approximately 60% appear to be pseudogenes [20]. Many alternative nomenclatures for hORs, including a comprehensive phylogenetic classification developed at the Weizmann Institute [17,22], have been proposed by various labs over the past few years.

The identification, cloning and sequence-based classification and analysis of candidate hORs are essential prerequisites for rational structure-function studies of this vast receptor family. Our goal was to identify the complete repertoire of hOR genes encoding full-length receptors. The approach was to carry out reiterative homology-based searches of GenBank DNA, particularly recently available unannotated raw sequences, and to compile hOR sequences already present in other public databases. We report here the identification and cloning of 347 putative full-length hOR receptor genes, which we believe accounts for nearly the entire repertoire of functional hORs. We also present a comparative sequence analysis of the predicted OR gene products and propose a new nomenclature for candidate hORs.

Results and discussion

Sequence database mining and odorant receptor cloning

The general strategy for the search for full-length hOR genes is shown in Figure 1. It was based on absence of introns in coding sequences of mammalian ORs [1,11] as well as high overall sequence similarity and the presence of several highly conserved sequence motifs in all known mammalian ORs [2].

The first step was to identify all currently known hOR sequences by extensive keyword and homology-based searches of several public DNA and protein sequence databases (see Materials and methods). The resulting several hundred sequences were compared with each other by BLAST and multiple sequence alignments. DNA and protein entries were matched. All duplicates were cross-referenced and apparent pseudogenes having frameshifts, deletions and other defects obviously incompatible with receptor function were eliminated. This initial screen identified approximately 90 *bona fide* full-length (according to criteria described below) hOR genes and a large number of candidate sequence fragments.

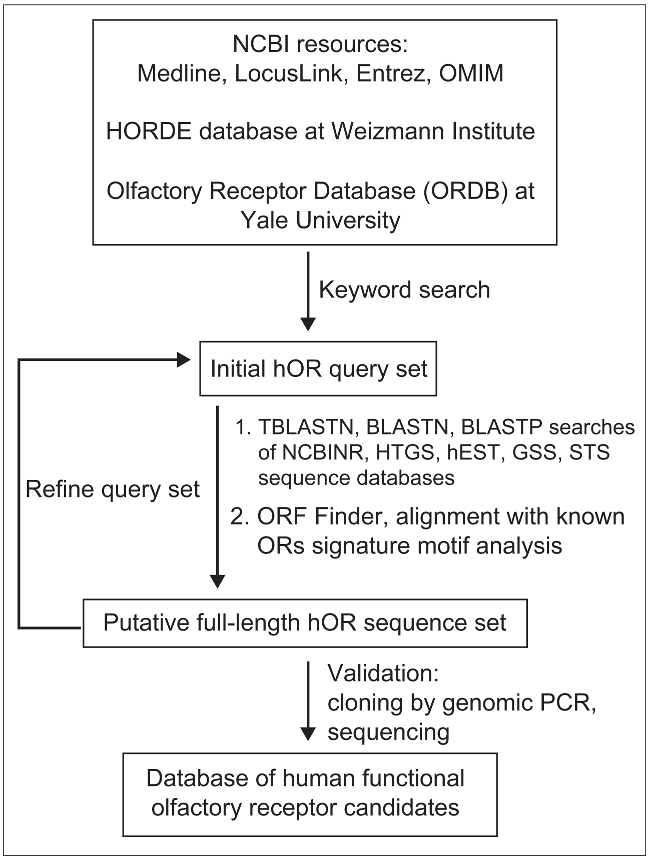
NCBI resources:
Medline, LocusLink, Entrez, OMIM
HORDE database at Weizmann Institute
Olfactory Receptor Database (ORDB) at Yale University
Keyword search
Initial hOR query set
1. TBLASTN, BLASTN, BLASTP searches of NCBINR, HTGS, hEST, GSS, STS sequence databases
2. ORF Finder, alignment with known ORs signature motif analysis
Refine query set
Putative full-length hOR sequence set
Validation:
cloning by genomic PCR,
sequencing
Database of human functional olfactory receptor candidates

Figure 1
Flow diagram for OR gene discovery by database mining.

The next step was identification of additional members of the family by exhaustive iterative homology searches of translated human genomic sequence databases, particularly the raw, unannotated high-throughput genomic sequences (HTGS), using protein queries corresponding to the known hORs. As the number of identified receptors grew, additional diverse sequences were added to the query list to capture additional ORs in databases. Genomic sequences containing areas of significant homology to ORs were subjected to open reading frame (ORF) searches. Several hundred identified ORFs of sufficient length (>250 amino acids) were translated and compared to known ORs. Three criteria were used to select putative ORs from this set: high overall sequence similarity, the presence of seven predicted transmembrane (TM) regions, and the presence of multiple positionally conserved sequence motifs, described in Materials and methods, that serve as signatures of this gene family. The basic criteria for recognizing a particular OR gene as encoding a full-length, functional receptor candidate were the presence of an uninterrupted ORF starting with an ATG codon and a complete seven-transmembrane unit. A few putative ORs were discarded from the final set because they contained structural features deemed to be incompatible with receptor function, such as insertions of multiple nucleotide sequence repeats in their ORFs.

These are intentionally minimalist criteria designed to exclude apparent OR pseudogenes, but not cryptic pseudogenes, such as those having defects on the level of transcription or splicing or containing subtle but functionally disruptive mutations in receptor coding region. For example, human OR17-24 (hOR17.01.01 in our nomenclature as discussed below) classified by others [23] as a probable pseudogene on the basis of deviations from other ORs in its carboxyl terminus and 5′-untranslated region (UTR) sequence is considered a regular candidate OR according to our analysis. It has to be noted that the final assignment of OR gene or pseudogene status must be based on functional expression data.

As a result of the search described above, 347 putative full-length OR genes have been identified in the human genome. This number includes all the previously known, annotated hOR sequences extracted from the public databases. It is feasible that a small number of full-length hORs escaped detection because of frameshifts or other sequencing or assembly errors in the corresponding HTGS entries. We are continuing routine searches using updated versions of genomic sequences to identify such cases. Because of the very high occurrence of OR pseudogenes in humans [15,16] and the presence of ORs in highly variable parts of human genome [13,24], it is also possible that some polymorphic members of this gene family exist in the human population in both intact and defective allelic forms. A small subset of identified OR ORFs was discarded because of such defects as partial deletions of TM1 or TM7 regions. The argument can be made that these OR genes could encode functional receptors. In addition, it was hypothesized that odorant reception may also be mediated by receptors of a completely different class, such as guanylate cyclases [25]. These caveats notwithstanding, we estimate that we have identified at least 90-95% of all full-length prototypical ORs in the human genome.

Cloning coding regions of full-length human olfactory receptor genes

To validate sequences extracted by genomic database mining, all hOR gene coding sequences described in this paper were cloned by direct genomic PCR using mixed DNA from ten individuals as a template and oligonucleotide primers corresponding to the amino- and carboxy-terminal sequences of the corresponding ORFs. An average of eight independent clones for each receptor were isolated and sequenced in their entirety. While single-nucleotide polymorphisms (SNPs) were detected in many hORs (data not shown), the sequencing data essentially confirmed correct identification of full-length OR-encoding ORFs. Sequences shown in the multiple sequence alignment (Figure 2) are those extracted from genomic sequence database entries.

Studies of human OR gene and pseudogene sequences covering approximately 150 full-length receptor genes [17,18] were recently published. A larger set of annotated putative human OR genes identified by an automated search algorithm became very recently accessible through a worldwide web interface [19]. A detailed comparison and cross-referencing of this overlapping set of data with the output of our independent database mining and genomic cloning effort allowed us to verify the hOR selection reported in this paper. The sets of the hORs identified by us and the other group strongly overlap, which corroborates the essentially complete overview of the hOR repertoire. We found differences between our data and those in the HORDE database, however. These include 29 hOR genes that are apparently identified as pseudogenes in the HORDE database, but encode functional hOR candidates by our analysis, as well as 10 hORs not found in HORDE (see Materials and methods). Our extensive cloning and sequencing data supported the conclusions presented here. Differences are probably caused by DNA sequencing errors in raw high-throughput genomic sequence. Other differences include single nucleotide/amino acid changes as well as discrepancies in the definition of amino and carboxyl termini for some ORs. An important caveat in interpreting these discrepancies is an apparent evolutionary decline in the human olfactory apparatus. Many specific anosmias (inabilities to smell a particular odorant) have been identified in humans [26,27]. Although the molecular basis for these anosmias is not currently known, all or some of them could be caused by hereditary OR defects. The high proportion of hOR pseudogenes [15,16] and the unusually high rate of SNPs in hORs [28] point to significant variability in the composition of the functional OR repertoire in the human population. More research is needed to clarify these issues and to generate the final catalog of functional hORs.

Genomic localization and general features of hORs

It has been previously demonstrated that members of the hOR gene family are distributed on all but a few human chromosomes. Through fluorescence *in situ* hybridization analysis, Rouquier [15] showed that OR sequences reside at more than 25 locations in the human genome and that the human genome has accumulated a striking number of dysfunctional copies: 72% of these sequences were found to be pseudogenes. We identified a total of 347 putative functional hOR genes located in multiple clusters on all human chromosomes, except for 2, 4, 18, 20, 21, and Y, with the majority (155 hORs) on chromosome 11. Chromosome 11 is followed in frequency by chromosome 1 (42 ORs), 9 (26 ORs) and 6 (24 ORs). By contrast, chromosomes 10, 22 and X seem to carry only a single full-length OR gene. Full-length hOR genes are present in more than 50 distinct clusters and are

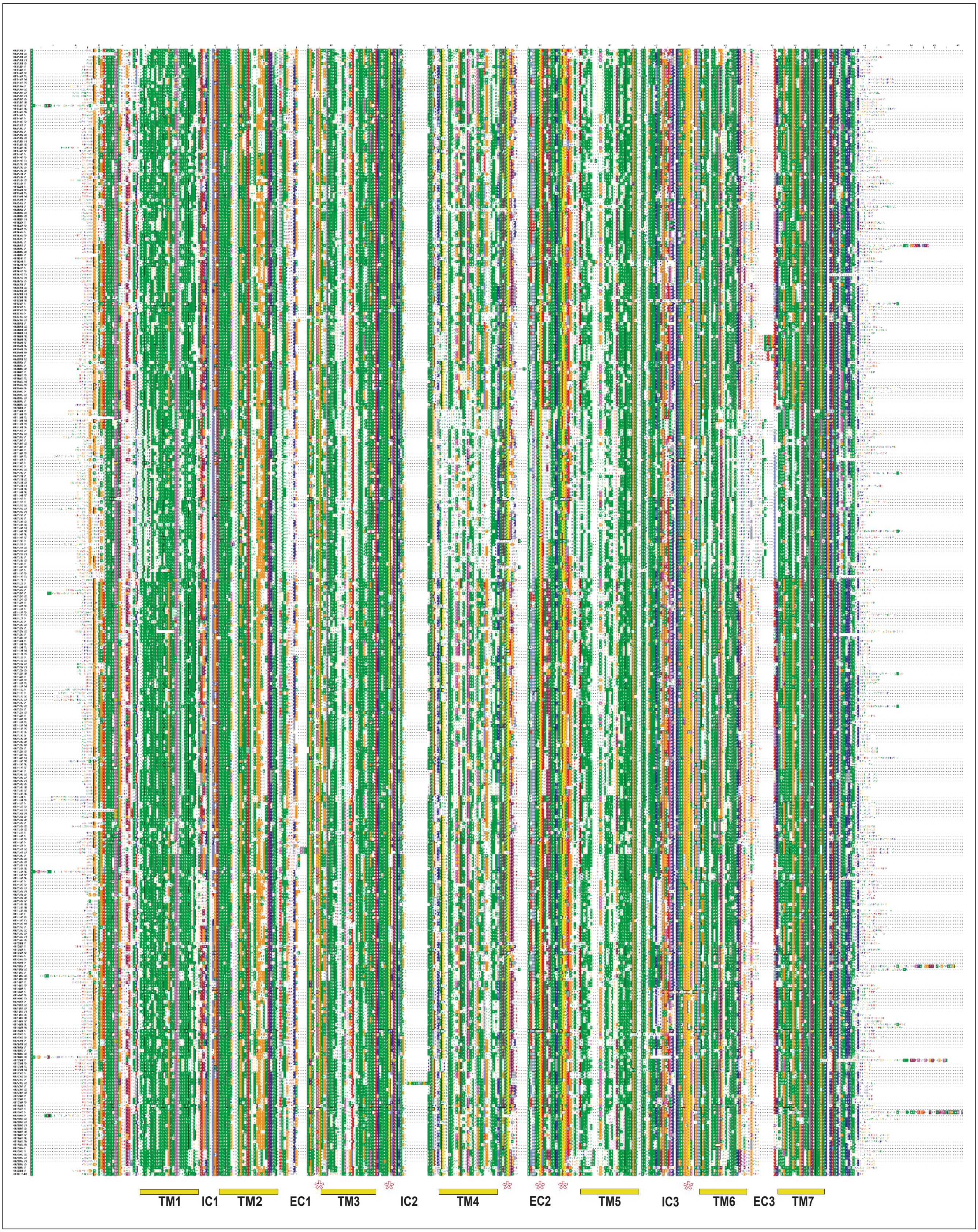

Figure 2 (see legend on next page)

Figure 2 *(see previous page)*
Protein sequence alignment of deduced amino-acid sequences of 347 human olfactory receptors. The alignment is color-coded according to the physicochemical properties and degree of conservation of amino-acid residues. The following side-chain-based coloring convention for groups of amino acids was used: Ile, Leu, Met, Val, Phe, Tyr, Ala, Trp (green), Lys, Arg (blue), His (light blue), Asp, Glu (red), Ser, Thr (orange), Pro (purple), Gly (magenta), Cys (yellow), Gln, Asn (grey). Positions with greater than 50% conservation (based on the amino-acid similarity groups shown above and PAM250 matrix) are shown as inverse characters on solid background. At the bottom of the figure, locations of the predicted membrane-spanning domains (TM1-7), intracellular loops (IC1-3) and extracellular loops (EC1-3) are indicated by horizontal bars. Positions of the three pairs of conserved cysteine residues are shown as asterisks. This figure is available as an additional data file in PDF format with the online version of this article.

interspersed with pseudogene OR sequences to a variable degree (data not shown).

The distribution of hORs from the two most abundant sources, chromosomes 11 and 1, in the phylogenetic tree of the OR repertoire is illustrated in Figure 3. While it does reveal some phylogenetic 'superclusters' of ORs, the chromosome-specific subsets of ORs are typically scattered across the phylogenetic tree. In addition to harboring almost half of the functional OR candidates, chromosome 11 contains an intriguing large cluster of 53 full-length OR genes (families 11.01 through 11.20 in Figure 2), which is evolutionarily distinct from the rest of the repertoire and is located near the telomeric end of chromosome 11p. Despite the phylogenetic uniqueness, this subgroup has strongly conserved sequence signature motifs of ORs. The remainder of ORs from chromosome 11 are clustered in the centromeric region (q11-12) or at the other end of the chromosome (q24-25). Although the available information on genomic localization of the OR genes was extracted from the Human Genome Project data and used in developing OR nomenclature (see below), mapping the exact chromosomal position of each OR gene or studying their cluster organization was not the goal of this study. Instead, our focus was on identification and comparative analysis of putative full-length hOR sequences produced by conceptual translation of the corresponding genomic OR ORFs.

Whereas the amino-acid identity between a few of the most distant human ORs is as low as 20%, the average identity for a random pair of hORs is in the 35-40% range. This lower limit of sequence identity within the hOR family is similar to that reported for candidate *Drosophila* OR genes [29] and chemosensory receptors of *Caenorhabditis elegans* [30]. An average predicted hOR is approximately 315 amino acids long, whereas the shortest OR included in the list of full-length receptors according to our criteria is 291 amino acids (hOR06.12.03). Uncertainty in determining exact amino termini for many ORs (see below) makes it difficult to identify the longest receptor sequence; however, one hOR (hOR19.03.01) is at least 355 amino acids long.

Strictly speaking, a candidate hOR would be the most appropriate designation for a protein belonging to this structurally distinct human G-protein-coupled receptor family, as their odorant-detecting function has not yet been reported. As pointed out in a recent review [3], very little experimental data exists on the expression of OR genes in human olfactory sensory neurons [12,31]. Some studies indicate that OR genes can be expressed in tissues other than the olfactory epithelium, suggesting potential alternative biological roles for this class of chemosensory receptors. Expression of various ORs was reported in human and murine erythroid cells [32], developing rat heart [33], avian notochord [34] and lingual epithelium [35]. The best experimentally documented case is the existence of a large subset of mammalian ORs transcribed in testes and expressed on the surface of mature spermatozoa, suggesting a possible role for ORs in

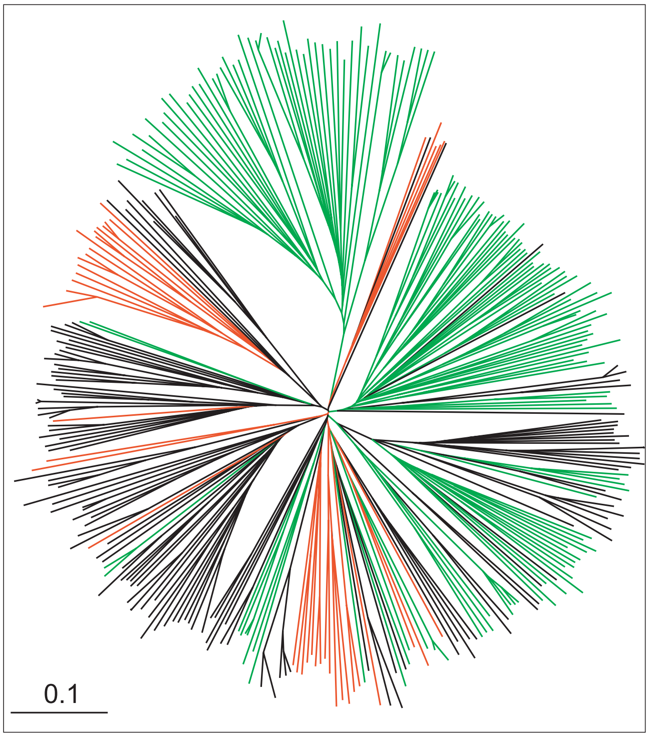

Figure 3
Unrooted phylogenetic tree of the human ORs. The tree was generated by ClustalX program using 347 hOR protein sequences. Positions of olfactory receptors residing on chromosome 11 (green) and 1 (red) are highlighted. The scale bar corresponds to the graphical distance equivalent to 10 % sequence divergence.

sperm chemotaxis [36-39]. It has been also hypothesized that olfactory receptors might provide molecular codes for cell-cell recognition in development and embryogenesis [40], including providing guidance for olfactory bulb glomeruli targeting by chemosensory neurons [9].

Conserved sequence motifs

Multiple sequence alignment (Figure 2) reveals single residues and more than a dozen amino-acid sequence motifs of various lengths conserved across the whole OR family, defining a sequence signature motif of an OR family member. Identification of each family member was confirmed by the signature motif conservation. Structural hallmarks of mammalian ORs based on comparison of smaller sets of receptors have been previously reviewed [2,3,5,41,42]. In addition to the traditional multiple sequence alignment (Figure 2), we used the 'sequence logo' presentation [43,44] to show the pattern of sequence conservation in all 347 full-length human ORs (Figure 4). Contrary to the traditional consensus sequences and multiple sequence alignments, this approach allows much more informative and easily interpretable visual representation of motifs and areas of significant sequence conservation in large protein families. Some conserved features are discussed in the following sections.

Transmembrane domains and loops

Strong sequence conservation is apparent in the intracellular loops, probably reflecting interactions of ORs with common intracellular partners, such as G_{olf} [45]. One such previously described conserved motif located at the junction of transmembrane domain (TM) 3 and intracellular loop (IC) 2 and incorporating E/DRY sequence conserved in all G-protein-coupled receptors is MAYDRYVAIC (single-letter amino-acid notation). The highly conserved motif KAFSTCXSH (X is any amino acid) in IC3 contains one of the cysteines from the intracellular pair conserved in many G-protein-coupled receptors, as well as serines and threonines potentially involved in phosphorylation events as discussed below. Other highly conserved regions include TM1, TM2 and TM7; the last two domains are routinely used in the design of family-specific degenerate oligonucleotide primers for PCR amplification of ORs. Most of the sequence variability is observed in extracellular loops EC1 and EC3, membrane-spanning domains TM4, TM5 and to a lesser degree TM3 and TM6, as well as in the extreme amino and carboxyl termini of the receptors (not shown in Figure 4). Structural diversity of these extracellular loops and transmembrane domains is believed to reflect ligand-binding specificity of ORs [1,42]. One model of a ligand-binding pocket in ORs stipulates the existence of 20 variable amino-acid residues on transmembrane helices 3, 4 and 5 that constitute the putative ligand 'complementarity-determining region' [5,42]. Taking into account the high observed sequence variability (Figures 2, 4) in additional OR domains known to be involved in ligand recognition in other G-protein-coupled receptors [46,47], such as TM6 and extracellular loops, that

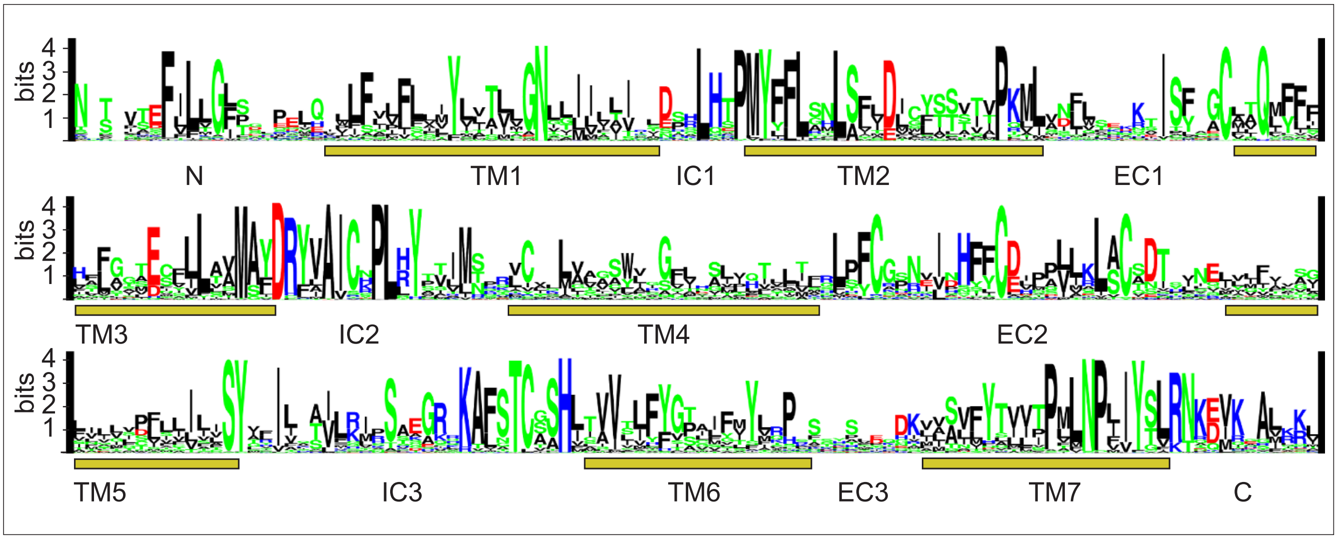

Figure 4
Conserved sequence motifs in the OR family. Sequence motifs based on analysis of the multiple sequence alignment of the full-length hOR repertoire are represented as 'sequence logo' [43,44]. The height of each amino-acid symbol in a given position is proportional to its frequency of occurrence. The height of the sequence information part is computed as the relative entropy between the observed fractions of a given amino-acid symbol and the respective *a priori* probabilities (assumed to be 0.05 for each amino acid). Locations of predicted transmembrane segments (TM1-7) as well as intracellular loops (IC1-3) and extracellular loops (EC1-3) are shown. TM boundaries are based on the predictions done using the neural network algorithm implemented with the PHDhtm and PHD topology programs [42,77,78]. The alignment was edited to adjust for OR length heterogeneity before sequence logo generation as described in Materials and methods. This figure is available as an additional data file in PDF format with the online version of this article.

model may not be sufficient to provide a complete description of odorant binding specificity determinants.

Almost all G-protein-coupled receptors contain a highly conserved cysteine in EC1 and another in EC2. There is direct and circumstantial evidence that the residues form an intra- or inter-molecular disulfide bond in many G-protein-coupled receptors, and in some cases these are critical for their ligand recognition and membrane trafficking [48-50]. OR sequences contain one highly conserved cysteine in EC1 and three in EC2, an unusual feature shared by only a few G-protein-coupled receptors, such as chemokine or P2Y1 receptors, which have an additional pair of disulfide-forming cysteines in the amino terminus and EC3 [50,51]. These four highly conserved cysteines suggest that two disulfide bonds may be present on the extracellular surface of ORs. Like other G-protein-coupled receptors, ORs also contain a conserved pair of cysteine residues in IC1 and IC2 (Figures 2,4).

Amino termini

Olfactory receptors have short (approximately 25-30 amino acids) extracellular amino termini that do not have homology to traditional cleavable leader peptides. A strongly conserved stretch of unknown functional significance, EF(I/L)LLG(L/F), is located approximately ten amino acids upstream of the predicted first transmembrane region. Like other G-protein-coupled receptors, ORs contain consensus sequences (NXS/T) for *N*-linked glycosylation near their amino termini and in EC1. Almost every OR has a single predicted amino-terminal *N*-glycosylation site and approximately 5% of the repertoire contains a consensus glycosylation site in EC1, whereas none are detected in EC2. A number of ORs also contain NXC motifs in the same locations. It was recently suggested that this motif, occurring in some murine ORs, is a possible novel *N*-linked glycosylation site [52]. It has to be noted that exact prediction of amino-terminal sequences of hORs based on genomic sequence information is not always clear-cut. While the OR coding sequences are generally believed to be intronless, there is at least one report of an exception to this rule resulting in an amino-terminal sequence heterogeneity [38]. In addition, 5′-UTR introns in OR genes are typically located very close to the OR coding region [53]. Extensive alternative splicing of 5′-UTRs of OR mRNAs have been described [38,54]. Therefore, both the straightforward conceptual translation of the longest OR ORF derived from genomic sequence as well as use of intron/exon prediction programs may result in incorrect identification of OR amino termini. The presence of a potential splice acceptor site, homology to other ORs, and an initiating ATG sequence context conducive to *in vivo* translation [55] were taken into consideration to predict amino-terminal OR protein sequences in our analysis.

Carboxyl termini

Predicted carboxyl termini of the ORs are short, with an average length of about 21 amino acids, and have significant similarity in their TM7-proximal half, starting with the highly conserved consensus RNK motif immediately adjacent to TM7. Many class I G-protein-coupled receptors, including rhodopsin and the β_2-adrenergic receptor, contain one or two cysteine residues in their carboxyl termini about 12 amino acids from the end of TM7. These correspond to a putative palmitoylation site which allows formation of a ‘fourth cytoplasmic loop’ that may be involved in G-protein coupling. Only 26% of the human ORs have one or more cysteines in their carboxy-terminal regions, suggesting that carboxy-terminal palmitoylation, if it occurs, is not general for ORs. An additional feature of OR carboxyl termini is a high content of positively charged amino-acid residues occurring in a positionally conserved pattern, a not uncommon feature of many other G-protein-coupled receptors. Yet another typical feature of many of these receptors is the presence of multiple serine and threonine residues in their carboxyl termini and IC3 that serve as phosphorylation sites for G-protein-coupled receptor kinases (GRKs) as well as protein kinases C (PKC) and A (PKA), a mechanism involved in agonist-dependent desensitization of the receptors. Mammalian ORs are known to undergo rapid desensitization after odorant stimulation [56] and GRK3 has been implicated in this process [57,58]. Approximately 20% of the ORs do not have any serine or threonine residues in their carboxyl termini, whereas the remaining 80% have an average of more than two such residues, most of which are located in the vicinity of positively charged amino acids and conform to consensus sequences for phosphorylation by PKC or PKA. In addition, the short third intracellular loop of all ORs contains four to five highly positionally conserved serine or threonine residues interspersed with four equally conserved basic amino acids, comprising a good potential site for phosphorylation.

A proposed nomenclature for functional human olfactory receptor candidates

By analogy with the classical examples of pharmacology-based nomenclatures of some well-studied G-protein-coupled receptor families, a nomenclature of functional olfactory receptors could be based on their odorant specificity. As such information is currently almost nonexistent for human ORs, other criteria, such as the chromosomal localization of their genes, similarly to the recently proposed *D. melanogaster* OR nomenclature [59], or their structural phylogenetic analysis, as proposed for mammalian ORs [17,22] should be used. We suggest an alternative hybrid nomenclature reflecting both phylogenetic clustering of OR gene products and their chromosomal localization. It is conceivable that phylogenetic closeness and high sequence similarity of ORs may reflect similarity in their ligand specificities. On the other hand, co-localization of a subset of OR genes in a particular genomic cluster might indicate their coordinate regulation and common biological function. One possible example of the latter is the OR gene cluster located on human chromosome 6 at the telomeric end of the HLA complex region. It has been shown recently that OR members from this cluster exhibit unusually high allelic variability, similar to the major histocompatibility complex (MHC) genes [24]. It is hypothesized that this group of human ORs could be functionally linked to MHC and involved in MHC-mediated mate preferences [60,61]. Another relevant example is the discovery of a block of human OR genes, at least one of which (OR7501 = hOR19.06.01) encodes a potentially functional receptor, in subtelomeric regions that are present in 7 to 11 similar copies on several chromosomes [13]. The presence of OR genes in these polymorphically multiplied, rearrangement-prone areas hints at a higher level complexity and individual variations in OR repertoire.

We generated phylogenetic trees using the neighbor-joining algorithm [62], as implemented in the two standard phylogenetic analysis software packages, ClustalX [63,64] and Phylip [65]. In both cases, the dataset was bootstrapped [66] to provide the statistical estimation of the reliability of the resulting tree topology. Two resulting dendrograms having very similar topology were used to identify OR clusters and assign the receptors to a particular family. A phylogenetic dendrogram derived from comparison of full-length OR protein sequences is shown in Figure 5, with the families bracketed. Strong phylogenetic clustering supported by the bootstrap tests (typically, only the bootstrap values higher than 50% were considered significant) as well as high sequence similarity (>40% identity) were used as a main criterion for grouping receptors into a particular family. The second criterion for defining a family was common chromosomal localization, including co-localization in a particular genomic locus if known at the time of analysis. It is believed that local tandem gene duplication is a common mechanism of OR evolution from common predecessors. Not surprisingly, there is a strong correlation between localization of an OR in a particular chromosomal cluster and its position in a phylogenetic dendrogram derived from comparison of full-length OR protein sequences (Figure 5). However, in a number of cases, ORs from different chromosomal loci converge in a phylogenetic cluster, implying recent gene duplication with interchromosomal insertion events from a distant region of the genome.

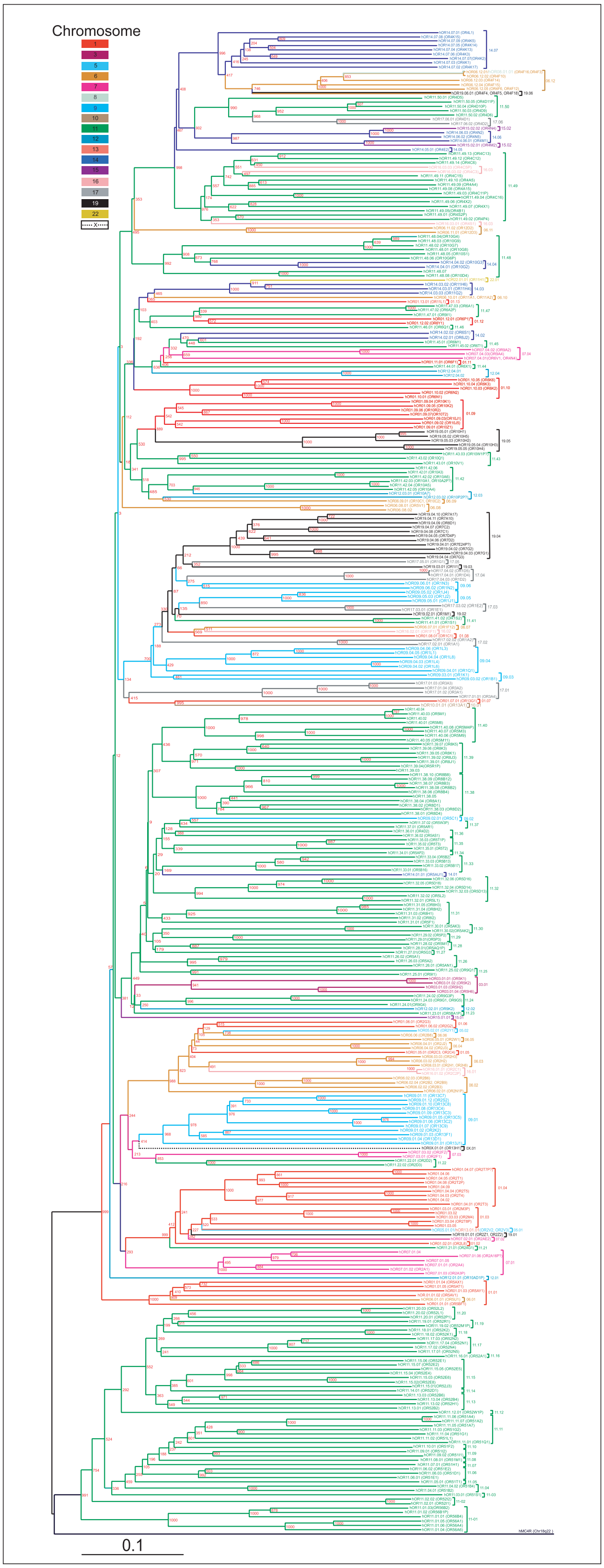

Figure 5

Figure 5
Rooted phylogenetic tree of the human ORs. The tree was generated by ClustalX program using 347 hOR protein sequences with the human melanocortin-4 receptor sequence as a root. Fragments starting from the conserved amino-terminal *N*-glycosylation site and ending 12 residues downstream of predicted TM7 were used for phylogenetic analysis. The numbers indicate the bootstrap value from 1,000 replicates. The bar corresponds to the graphical distance equivalent to 10% sequence divergence. OR names in parentheses correspond to HORDE analogs of the OR sequences identified by us. This figure is available as an additional data file in PDF format with the online version of this article.

As a result of this branching analysis, the family of 347 hORs was subdivided into 119 families from 1 to 14 members in each (Figure 5). The minimum pairwise intrafamily amino-acid identity is 43%, and the average minimum amino-acid identity within all of the 77 families having more than one member is 62%. Each chromosome-specific subset of hORs consists of 1 to 50 (for chromosome 11) independently numbered families, whereby each chromosome has family 1, for example. Members of each family are numbered sequentially from 1. An example of our nomenclature is hOR11.01.02, which represents a human olfactory receptor located on chromosome 11, from family 1 and which is member 2 of the family.

Numerous ways of classifying ORs on the basis of cloning technique, clone name or genomic location have been put forward. The most advanced and consistent existing nomenclature, which includes both genes and pseudogenes and is implemented in the HORDE database, is based solely on phylogenetic clustering with consequent division of ORs into families and subfamilies [17,22]. According to this nomenclature, receptor sequences with 40% or more amino-acid identity are considered members of the same family, whereas those sharing 60% or more identity constitute a subfamily. We believe that, at least in some cases, OR names in our nomenclature carry more biologically relevant information and could be more rational. Consider, for example, receptors hOR22.01.01 (hOR11H1), hOR14.03.01 (hOR11H4), hOR14.03.02 (hOR11H6) and hOR14.03.03 (hOR11G2), where the HORDE names are in parentheses. According to our nomenclature (see Figure 5) these four receptors form two complete families, 22.01 with one member located on chromosome 22, and 14.03 with three chromosome 14 members located in a single gene cluster. According to the HORDE division, all four receptors belong to the family 11. Three ORs, including the one from chromosome 22 and two from chromosome 14, fall into subfamily H, whereas the remaining chromosome 14 OR belongs to subfamily G. In other cases, similar OR grouping is provided by both nomenclatures. Nevertheless, even for those cases, the names in our nomenclature carry information about the chromosome of origin for a given OR instead of an arbitrary (sub)family name. An added convenience of the nomenclature we are proposing is the straightforward, 'computer-friendly' format, which would facilitate the handling of hundreds of OR sequences. It allows encoding up to 99 OR subfamilies of 99 members for each chromosome, possibly including polymorphic OR forms, as well as easy sorting of receptor lists.

For the reasons discussed earlier, some additional functional hOR candidates are likely to be identified in the future, and will have to be accommodated by the nomenclature. It may also need to be refined as we learn more about the biological function and specificity of olfactory receptors and the detailed genomic mapping of OR clusters. We believe, however, that the OR nomenclature we describe represents a comprehensive, convenient and possibly more biologically relevant alternative to the existing OR classifications.

Conclusions

Identification and cloning of the functional hOR repertoire creates a basis for addressing many unresolved issues in human olfaction. Most importantly, in conjunction with robust heterologous expression and assay systems and high-throughput screening of odorant libraries, it will ultimately lead to understanding of structure-function relationships and small-molecule recognition by this large group of G-protein-coupled receptors. The impact of genetic polymorphism of ORs on differential olfactory perception in the human population is another exciting topic. Global comparative analysis of functional hOR candidate gene and pseudogene repertoires, as well as of repertoires of human and murine ORs will shed light on the evolution of the human olfactory apparatus and its biological consequences.

Materials and methods

Sequence database mining

Human OR gene sequences were extracted from the following online public sequence databases: National Center for Biotechnology Information (NCBI) protein and DNA database resources [67] (queried via several interfaces including LocusLink, UniGene, Entrez and BLAST); the Human Olfactory Receptor Data Exploratorium (HORDE) at the Weizmann Institute, Israel [17,19]; and the Olfactory Receptor Database (ORDB) at Yale University [41,68,69].

Similarity-based searches for putative hOR gene sequences were performed by querying NR, HTGS, EST, GSS and STS divisions of GenBank using the tblastn and blastp search algorithms [70]. The blastp searches were run using default parameters with the expect value cut-off of 1×10^{-8}.

ORFs in DNA sequences were identified and analyzed using the GeneQuest program (LaserGene, DNASTAR). All ORFs longer than 250 amino acids were conceptually translated and compared to the entirety of GenBank by the tblastn and blastp searches. The presence of no fewer than five of the following consensus sequence signature motifs (see Figure 4) was used as a criterion for recognizing ORs: EF(I/L)LLG(L/F) (upstream of TM1), PMYFFL (TM2), MA(Y/F)DRY(V/L/A)AIC (junction of TM3 and IC2), three C (EC2), SY (junction of TM5 and IC3), KAFSTCXSH (IC3), P(M/L)LNP(L/I/F)IY (TM7).

Functional hOR candidates identified by us were compared with the hOR sequences in the HORDE database using blastp and blastn searches. HORDE counterparts for the following hORs identified by us are missing: 06.08.02, 07.01.04, 07.01.05, 11.38.05, 11.40.04, 11.42.06, 11.48.07, 12.04.01, 12.04.02, 15.01.01. The following functional hORs candidates are apparently identified as pseudogenes in HORDE: 01.03.01 (OR2M3P), 01.03.04 (OR2T8P), 01.04.07 (OR2T7P), 01.04.08 (OR2T2P), 06.02.01 (OR2N1P), 07.01.03 (OR2A3P), 07.01.06 (OR2A16P), 11.01.02 (OR56B1P), 11.12.01 (OR52W1P), 11.19.02 (OR52M1P), 11.23.01 (OR5BA1P), 11.24.02 (OR9G3P), 11.28.01 (OR5AQ1P), 11.35.03 (OR5T1P), 11.37.02 (OR5W3P), 11.39.04 (OR5R1P), 11.40.08 (OR5M4P), 11.42.03 (OR10A2P), 11.43.03 (OR10W1P), 11.47.02 (OR6A2P), 11.49.01 (OR4S2P), 11.49.03 (OR4C11P), 11.50.05 (OR4D11P), 12.01.01 (OR10AD1P), 12.03.02 (OR10P2P), 16.01.02 (OR2C2P), 16.03.03 (OR4C5P), 19.04.01 (OR7E24P), 19.04.05 (OR7D4P).

Sequence alignments and phylogenetic analysis

Multiple sequence alignments were constructed with ClustalX [63,64]. The alignments were slightly modified to adjust the gap positions by visual inspection. Alignment editing and shading was done using GeneDoc Multiple Sequence Alignment Editor [71] and BioEdit Sequence Alignment Editor [72]. Phylogenetic analysis was done using either ClustalX or Phylip 3.6 [65] software using the neighbor-joining method [62] and standard parameters. The resulting datasets were subjected to 1,000 (Clustal) or 100 (Phylip) rounds of bootstrapping [66] to evaluate the statistical significance of the consensus tree topologies. To root the phylogenetic trees, the human melanocortin receptor-4 sequence was used as an outgroup. Tree representations were constructed using TreeView [73] and TreeExplorer (K. Tamura, personal communication).

Sequence logos of hORs were generated using a web-based program developed by Jan Gorodkin [74]. For most analyses, short parts of the receptor sequences corresponding to the extreme amino termini (upstream of the conserved predicted *N*-linked glycosylation site) and carboxyl termini (beyond 12 amino acids residues downstream of TM7) were deleted to avoid length heterogeneity in the compared sequences. No significant sequence conservation was observed in these short removed receptor fragments of variable lengths. Similarly, all alignment gap positions where more than 95% of all ORs had a gap were deleted before the sequence logo generation. The SpliceView intron/exon boundary prediction program [75,76] was used to identify potential splice sites in predicted 5′-UTRs of OR mRNAs. Positions of transmembrane domains in ORs were predicted using the PHD topology routine [77] at PredictProtein server at EMBL, Heidelberg [78].

Genomic PCR and cloning

Human genomic DNA was amplified by polymerase chain reaction (PCR) usinq Taq polymerase (HotStarTaq, Qiagen) and oligonucleotide primers based on sequences corresponding to the extreme amino- and carboxy-terminal parts of OR ORFs. A commercial mixture of genomic DNA from 10 unknown individuals (5 male, 5 female) was used (Novagen) at 50 ng per 50 µl PCR reaction. PCR was run at 50 µM final oligonucleotide primer concentrations and 0.25 mM concentration of each dNTP. The cycling conditions were as follows: initial denaturation at 94°C for 15 min, denaturation at 94°C for 40 sec, annealing at 50-60°C for 40 sec, and extension at 72°C for 90 sec, for 35 cycles, followed by 7 min extension at 72°C. Intrinsic PCR error rate in these conditions was estimated in pilot experiments and was found to be not higher than 1 mutation per 2,500 bp of sequence. Despite the high sequence similarity of many OR genes and pseudogenes, no chimeric products of genomic PCR [4] were observed in the conditions used. PCR fragments were subcloned into the plasmid pcDNA3.1-V5/His (Invitrogen, TA kit), *Escherichia coli* DH5α cells were transformed, miniprep DNA was prepared using QIAPrep 96 Turbo Miniprep Kit (Qiagen) and subject to sequence analysis with vector-based primers using automated fluorescent sequencer ABI377 (PE Applied Biosytems). An average of eight individual clones were analyzed for every OR cloned. Sequence assembly and editing was done using SeqMan program (DNASTAR). The resulting sequences were compared with the original database sequence used to design the primers.

Acknowledgements

We thank Lubert Stryer, Mark Zoller, Peter Mombaerts, Paul Feinstein, Elliot Adler, Denis Baylor, Catherine Dulac and Arcady Mushegian for critical comments.

Additional data files

A FASTA file with the full-length sequences of the 347 hORs as well as PDF versions of figure 2, figure 4 and figure 5 are included in the online version of this article.

References

1. Buck L, Axel R: **A novel multigene family may encode odorant receptors: a molecular basis for odor recognition.** *Cell* 1991, **65:**175-187.
2. Mombaerts P: **Seven-transmembrane proteins as odorant and chemosensory receptors.** *Science* 1999, **286:**707-711.
3. Mombaerts P: **Odorant receptor genes in humans.** *Curr Opin Genet Dev* 1999, **9:**315-320.
4. Mombaerts P: **Molecular biology of odorant receptors in vertebrates.** *Annu Rev Neurosci* 1999, **22:**487-509.
5. Pilpel Y, Sosinsky A, Lancet D: **Molecular biology of olfactory receptors.** *Essays Biochem* 1998, **33:**93-104.
6. Dryer L, Berghard A: **Odorant receptors: a plethora of G-protein coupled receptors.** *Trends Pharmacol Sci* 1999, **20:**413-417.
7. Bockaert J, Pin JP: **Molecular tinkering of G protein-coupled receptors: an evolutionary success.** *EMBO J* 1999, **18:**1723-1729.
8. Chess A, Simon I, Cedar H, Axel R: **Allelic inactivation regulates olfactory receptor gene expression.** *Cell* 1994, **78:**823-834.
9. Mombaerts P, Wang F, Dulac C, Chao SK, Nemes A, Mendesohn M, Edmondson J, Axel R: **Visualizing an olfactory sensory map.** *Cell* 1996, **87:**675-686.
10. Malnic B, Hirono J, Sato T, Buck LB: **Combinatorial receptor codes for odors.** *Cell* 1999, **96:**713-723.
11. Nef P, Hermans-Borgmeyer I, Artieres-Pin H, Beasley L, Dionne VE, Heinemann SF: **Spatial pattern of receptor expression in the olfactory epithelium.** *Proc Natl Acad Sci USA* 1992, **89:**8948-8952.

12. Ben-Arie N, Lancet D, Taylor C, Khen M, Walker N, Ledbetter DH, Carrozzo R, Patel K, Sheer D, Lehrach H, North MA: **Olfactory receptor gene cluster on human chromosome 17: Possible duplication of an ancestral receptor repertoire.** *Hum Mol Genet* 1994, **3:**229-235.
13. Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, Collins C, Giorgi D, Iadonato S, Johnson F, *et al.*: **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* 1998, **7:**13-26.
14. Reed RR: **Mechanisms of sensitivity and specificity in olfaction.** *Cold Spring Harbor Symp Quant Biol* 1992, **57:**501-504.
15. Rouquier S, Taviaux S, Trask BJ, Brand-Arpon V, van den Engh G, Demaille J, Giorgi D: **Distribution of olfactory receptor genes in the human genome.** Nature Genet *1998,* **18:**243-250.
16. Rouquier S, Blancher A, Giorgi D: **The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates.** *Proc Natl Acad Sci USA* 2000, **97:**2870-2874.
17. Glusman G, Bahar A, Sharon D, Pilpel Y, White J, Lancet D: **The olfactory receptor gene superfamily: data mining, classification, and nomenclature.** Mamm Genome 2000, **11:**1016-1023.
18. Fuchs T, Glusman G, Horn-Saban S, Lancet D, Pilpel Y: **The human olfactory subgenome: from sequence to structure and evolution.** *Hum Genet* 2001, **108:**1-13.
19. **Human Olfactory Receptor Data Exploratorium (HORDE)** [http://bioinformatics.weizmann.ac.il/HORDE/]
20. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
21. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The Sequence of the Human Genome.** *Science* 2001, **291:**1304-1351.
22. Lancet D, Ben-Arie N: **Olfactory receptors.** *Curr Biol* 1993, **3:**668-674.
23. Glusman G, Clifton S, Roe B, Lancet D: **Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity.** *Genomics* 1996, **37:**147-160.
24. Ehlers A, Beck S, Forbes SA, Trowsdale J, Volz A, Younger R, Ziegler A: **MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes.** *Genome Res* 2000, **10:**1968-1978.
25. Gibson AD, Garbers DL: **Guanylyl cyclases as a family of putative odorant receptors.** Annu Rev Neurosci 2000, **23:**417-439.
26. Amoore JE: **Specific anosmia and the concept of primary odors.** *Chem Senses Flavor* 1977, **2:**267-281.
27. Amoore JE, Steinle S: **A graphic history of anosmia.** *Chem Senses* 1991, **3:**331-351.
28. Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, Sharon D: **Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor gene and pseudogenes.** *Nat Genet* 2000, **26:**221-224.
29. Vosshal LB: **Olfaction in *Drosophila*.** *Curr Opin Neurobiol* 2000, **10:**498-503.
30. Troemel ER, Chou JH, Dwyer ND, Colbert HA, Bargmann CI: **Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*.** *Cell* 1995, **83:**207-218.
31. Sosinsky A, Glusman G, Lancet D: **The genomic structure of human olfactory receptor genes.** *Genomics* 2000, **70:**49-61.
32. Feingold EA, Penny LA, Nienhuis AW, Forget BG: **An olfactory receptor gene is located in the extended human beta-globin gene cluster and is expressed in erythroid cells.** *Genomics* 1999, **61:**15-23.
33. Drutel G, Arrang JM, Diaz J, Wisnewsky C, Schwartz K, Schwartz JC: **Cloning of OL1, a putative olfactory receptor and its expression in the developing rat heart.** *Receptors Channels* 1995, **3:**33-40.
34. Nef S, Nef P: **Olfaction: transient expression of a putative odorant receptor in the avian notochord.** *Proc Natl Acad Sci USA* 1997, **94:**4766-4771.
35. Abe K, Kusakabe K, Tanemura Y, Emori Y, Arai S: **Multiple genes for G protein-coupled receptors and their expression in lingual epithelia.** *FEBS Lett* 1993, **316:**253-256.
36. Parmentier M, Liebert F, Schurmans S, Schiffmann S, Lefort A: **Expression of members of the putative olfactory receptor gene family in mammalian germ cells.** *Nature* 1992, **355:**453-455.
37. Walensky LD, Roskams J, Lefkowitz RJ, Snyder SH, Ronnett GV: **Odorant receptors and desensitization proteins colocalize in mammalian sperm.** *Mol Med* 1995, **1:**130-141.
38. Walensky LD, Ruat M, Bakin RE, Blackshaw S, Ronnett GV, Snyder SH: **Two novel odorant receptor families expressed in spermatids undergo 5′-splicing.** *J Biol Chem* 1998, **273:**9378-9387.
39. Branscomb A, Seger J, White RL: **Evolution of odorant receptors expressed in mammalian testes.** *Genetics* 2000, **156:**785-797.
40. Dreyer WJ: **The area code hypothesis revisited: Olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos.** *Proc Natl Acad Sci USA* 1998, **95:**9072-9077.
41. Skoufos E, Healy MD, Singer MS, Nadkarni PM, Miller PL, Shepherd G: **Olfactory receptor database: a database of the largest eukaryotic gene family.** *Nucleic Acids Res* 1999, **27:**343-345.
42. Pilpel Y, Lancet D: **The variable and conserved interfaces of modeled olfactory receptor proteins.** *Prot Sci* 1999, **8:**969-977.
43. Gorodkin J, Heyer LJ, Brunak S, Stormo GD: **Displaying the information contents of structural RNA alignments: the structure logos.** *Comput Appl Biosci* 1997, **13:**583-586.
44. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.
45. Jones DT, Reed RR: **Golf: an olfactory neuron specific-G protein involved in odorant signal transduction.** Science 1989, **244:**790-795.
46. Baldwin JM: **Structure and function of receptors coupled to G proteins.** *Curr Opin Cell Biol* 1994, **6:**180-190.
47. Flower DR: **Modelling G-protein-coupled receptors for drug design.** *Biochim Biophys Acta* 1999, **1422:**207-234.
48. Le Gouill C, Parent JL, Rola-Pleszczynski M, Stankova J: **Role of the Cys90, Cys95 and Cys173 residues in the structure and function of the human platelet-activating factor receptor.** *FEBS Lett* 1997, **402:**203-208.
49. Zeng FY, Wess J: **Identification and molecular characterization of m3 muscarinic receptor dimers.** *J Biol Chem* 1999, **274:**19487-19497.
50. Blanpain C, Lee B, Vakili J, Doranz BJ, Govaerts C, Migeotte I, Sharron M, Dupriez V, Vassart G, Doms RW, Parmentier M: **Extracellular cysteines of CCR5 are required for chemokine binding, but dispensable for HIV-1 coreceptor activity.** *J Biol Chem* 1999, **274:**18902-18908.
51. Hoffmann C, Moro S, Nicholas RA, Harden TK, Jacobson KA: **The role of amino acids in extracellular loops of the human P2Y1 receptor in surface expression and activation processes.** *J Biol Chem* 1999, **274:**14639-14647.
52. Xie SY, Feinstein P, Mombaerts P: **Characterization of a cluster comprising approximately 100 odorant receptor genes in mouse.** *Mamm Genome* 2000, **11:**1070-1078.
53. Glusman G, Sosinsky A, Ben-Asher E, Avidan N, Sonkin D, Bahar A, Rosenthal A, Clifton S, Roe B, Ferraz C, *et al.*: **Sequence, structure and evolution of a complete human olfactory receptor gene cluster.** *Genomics* 2000, **63:**227-245.
54. Bulger M, Bender MA, van Doorninck JH, Wertman B, Farrell CM, Felsenfeld G, Groudine M, Hardison R: **Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters.** *Proc Natl Acad Sci USA* 2000, **97:**14560-14565.
55. Kozak M: **The scanning model for translation: an update.** *J Cell Biol* 1989, **108:**229-241.
56. Zufall F, Leinders-Zufall T: **The cellular and molecular basis of odor adaptation.** *Chem Senses* 2000, **25:**473-481.
57. Boekhoff I, Inglese J, Schleicher S, Koch WJ, Lefkowitz RJ, Breer H: **Olfactory desensitization requires membrane targeting of receptor kinase mediated by beta gamma-subunits of heterotrimeric G proteins.** *J Biol Chem* 1994, **269:**37-40.
58. Peppel K, Boekhoff I, McDonald P, Breer H, Caron MG, Lefkowitz RJ: **G protein-coupled receptor kinase 3 (GRK3) gene disruption leads to loss of odorant receptor desensitization.** *J Biol Chem* 1997, **272:**25425-25428.
59. Drosophila Receptor Nomenclature Committee: **Unified nomenclature system for the *Drosophila* odorant receptors.** *Cell* 2000, **102:**145-146.
60. Fan W, Cai W, Parimoo S, Lennon GG, Weismann SM: **Identification of seven new human MHC class I region genes around the HLA-F locus.** *Immunogenetics* 1996, **44:**97-103.

61. Ziegler A, Ehlers A, Forbes S, Trowsdale J, Uchanska-Ziegler B, Volz A, Younger R, Beck, S: **Polymorphic olfactory receptor genes and HLA loci constitute extended haplotypes.** In *Major Histocompatibility Complex – Evolution, Structure and Function.* Edited by M Kasahara. Tokyo: Springer Verlag; 2000, 110-130.
62. Saitou N, Nei M: **The neighbour-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
63. Thomson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
64. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ: **Multiple sequence alignment with Clustal X.** *Trends Biochem Sci* 1998, **23:**403-405.
65. Felsenstein J: **PHYLIP – phylogeny inference package (version 3.2).** *Cladistics* 1989, **5:**164-166.
66. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39:**783-791.
67. **National Center for Biotechnology Information (NCBI)** [http://www.ncbi.nlm.nih.gov/]
68. **The Olfactory Receptor Database (ORDB)** [http://ycmi.med.yale.edu/senselab/ordb/]
69. Skoufos E, Marenco L, Nadkarni PM, Miller PL, Shepherd GM: **Olfactory receptor database: a sensory chemoreceptor resource.** *Nucleic Acids Res* 2000, **28:**341-345.
70. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/blast]
71. **GeneDoc Multiple Sequence Alignment Editor** [http://www.psc.edu/biomed/genedoc]
72. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41:**95-98.
73. Page RDM: **TreeView: An application to display phylogenetic trees on personal computers.** *Comput Appl Biosci* 1996, **12:**357-358.
74. **Protein Sequence Logos** [http://www.cbs.dtu.dk/gorodkin/appl/plogo.html]
75. **WebGene** [http://www.itba.mi.cnr.it/webgene/]
76. Rogozin IB, Milanesi L**: Analysis of donor splice signals in different organisms.** *J Mol Evol* 1997, **45:**50-59.
77. *Rost B: **PHD: predicting one-dimensional protein structure by profile based neural networks.*** Meth Enzymol *1996, **266:**525-539.*
78. **The PredictProtein server** [http://www.embl-heidelberg.de/predictprotein/predictprotein.html]