# The olfactory receptor gene superfamily of the mouse

Xinmin Zhang and Stuart Firestein

*Department of Biological Sciences, Columbia University, New York, New York, USA*
*Correspondence should be addressed to S.F. (sjf24@columbia.edu)*

Published online: 22 January 2002, DOI: 10.1038/nn800

**Olfactory receptor (OR) genes are the largest gene superfamily in vertebrates. We have identified the mouse OR genes from the nearly complete Celera mouse genome by a comprehensive data mining strategy. We found 1,296 mouse OR genes (including ~20% pseudogenes), which can be classified into 228 families. OR genes are distributed in 27 clusters on all mouse chromosomes except 12 and Y. One OR gene cluster matches a known locus mediating a specific anosmia, indicating the anosmia may be due directly to the loss of receptors. A large number of apparently functional 'fish-like' Class I OR genes in the mouse genome may have important roles in mammalian olfaction. Human ORs cover a similar 'receptor space' as the mouse ORs, suggesting that the human olfactory system has retained the ability to recognize a broad spectrum of chemicals even though humans have lost nearly two-thirds of the OR genes as compared to mice.**



The detection of environmental chemicals is mediated by peripheral olfactory organs of varied complexity in almost all metazoans. Typically, specialized sensory neurons initiate perception by detecting ambient molecules, commonly called odors, that interact with protein receptors in their membranes. The odor receptors (ORs), the molecular receptors that recognize odorant molecules, belong to the superfamily of seven-transmembrane-domain, G protein–coupled receptors (GPCRs). A family of receptor guanylyl cyclases have been proposed as receptors for odors or pheromones[1], but currently there is no functional data supporting this notion. Consequently, in this article 'ORs' refers to those receptors belonging to the GPCR superfamily.

Since their initial discovery in rat[2], ORs have been identified in various species of both invertebrates (nematode and fruit fly) and vertebrates (fish, amphibians, lizards, birds and mammals)[3]. Invariably the genes encoding ORs constitute a large gene family, and in mammals they constitute the largest gene superfamily in the genome. It has been estimated that there are ~1,000 ORs in the mouse and rat, ~500–750 in human and ~100 in fish[4,5]. In several cases these estimates, which were based largely on hybridization experiments, can now be further assessed by investigating sequenced genomes.

Handling data from such a large family of genes is not trivial, and until recently sequence information for ORs has been fragmentary in all mammalian species. Most sequences have been obtained from either cDNA or genomic DNA (most OR genes seem to be intronless) with a variety of PCR approaches. These have produced mostly gene fragments, with a smaller number of full-length genes. Among the drawbacks of these approaches are probable primer bias and possible recombination among highly related sequences, which may lead to many genes not being detected[6,7]. As a result, more than a decade after the first cloning of an OR, there are still relatively few gene sequences available from this family.

With some mammalian genomes now sequenced, analytical approaches to obtaining OR sequences have become feasible. After the release of the human genome, the human OR repertoire was thoroughly explored at the level of the whole genome[8,9]. About 900 OR genes, distributed in 24 clusters throughout the genome (except chromosomes 20 and Y), were discovered; however, 60% of these seemed to be pseudogenes and fewer than 350 were intact OR genes. This massive degeneration of ORs might be due to the lesser importance of olfaction in humans as compared to the visual and auditory senses. In contrast, mice are thought to have a much larger and more complex repertoire of ORs. Mice have become the preferred experimental animals in studies of olfaction, largely because of the success of gene targeting. Obtaining the complete mouse OR repertoire would therefore be invaluable. Until recently, only ~100 full-length mouse OR sequences were available in the public databases, more than half obtained by sequencing genomic regions of known OR gene clusters[10–12]. Nonetheless, a whole-genomic approach remains the most efficient and thorough means to retrieve the complete OR repertoire.

The Celera mouse genome was released in May 2000. Since then we have carried out a comprehensive data mining effort on the nearly complete mouse genome and found nearly the entire mouse OR gene repertoire, comprising roughly 1,300 OR genes. As expected, there are about 1,000 potentially functional genes (with the remaining ~20% probably pseudogenes) distributed in clusters throughout the chromosomes. We categorized 1,130 OR genes into 27 clusters, with an average frequency of 29 Kb per OR gene within clusters and with occasional intervening unrelated genes. This repertoire covers most OR genes currently in the public databases and provides full-length sequences for the ~200 fragments in various other databases. Future study of ORs will also benefit from this complete OR gene database, as full-length sequences can now be easily obtained by comparing

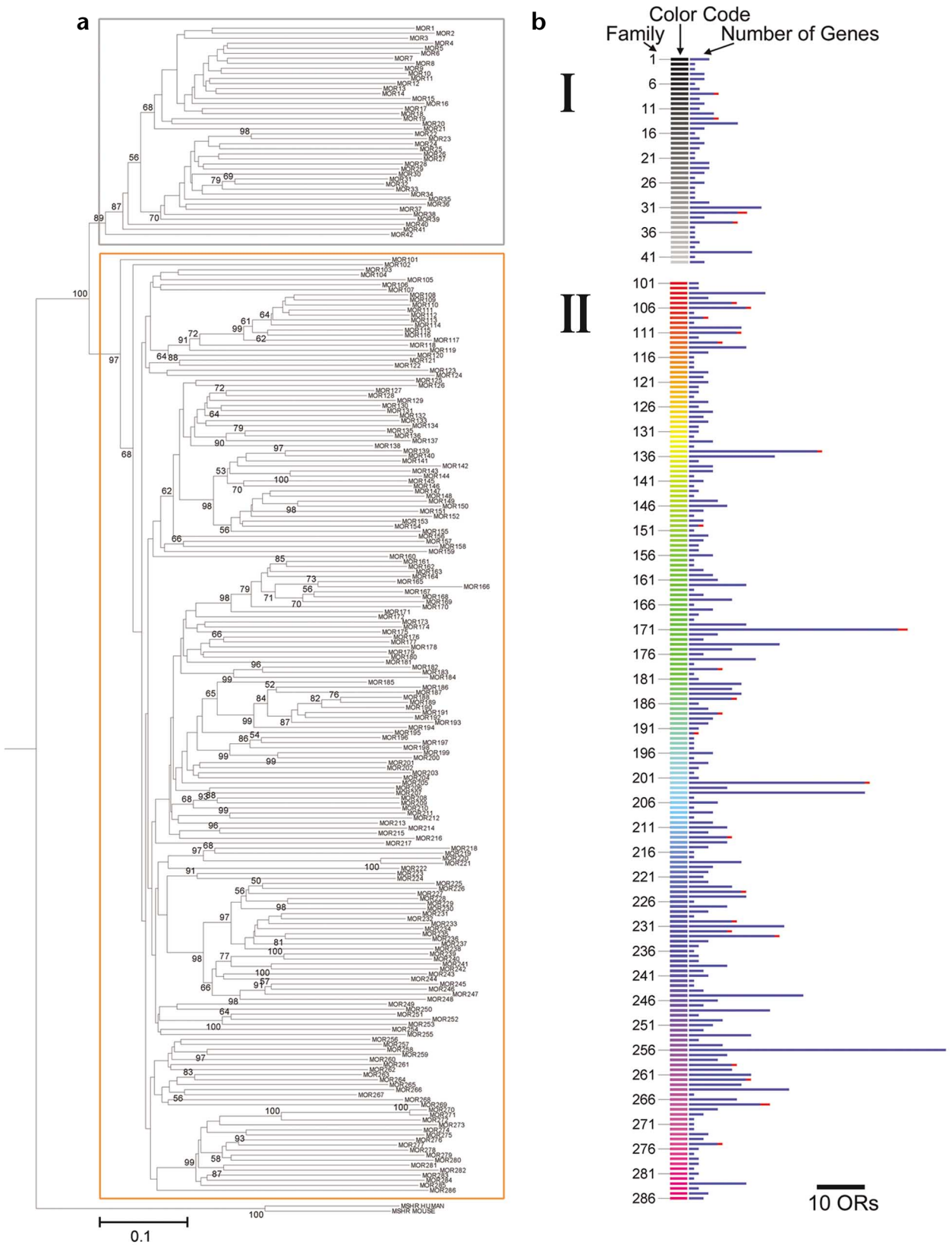


**Fig. 1.** Mouse OR families. (**a**) Phylogenetic tree of the consensus protein sequences of mouse OR families. Bootstrap values are shown at nodes with >50% support. The tree was rooted using human and mouse melanocyte-stimulating hormone receptors (MSH-R). Class I families are contained by the gray box and Class II families by the orange box. (**b**) List of all the families in the same order as in (**a**) (1–41 are Class I families, 101–286 Class II). Each family has been given a color code (different gray scale shades for Class I families and different color shades for Class II families). The number of intact genes (blue) and pseudogenes (red) in each family is shown after the color code.

**Table 1. Protein sequence comparison of genome-derived database with mouse OR genes in public database.**

| Database | OR genes from GenBank | OR genes from ORDB |
|---|---|---|
| Number of OR genes | 362 | 122 |
| Matches (>95% identity) | 327 (90.3%) | 110 (90.2%) |
| Matches (>99% identity) | 280 (77.3%) | 93 (76.2%) |

PCR fragments with the database, avoiding time-consuming RNA RACE or genomic library screening.

## Results

### The mouse OR repertoire

We have developed a comprehensive system to search for candidate OR gene sequences in the Celera Assembled and Annotated Mouse Genome. Briefly, TBLASTN searches for OR sequences were instituted using known human and mouse OR sequences as queries. The output sequences were subjected to further analysis incorporating ORF discovery, profile HMM searches and BLASTP searches to determine which were true OR sequences. Exhaustive TBLASTN searching was continued until no new output sequences were found. A conceptual translation method using FASTY3 and a database comprising ~1,000 full-length mammalian ORs was used to determine the coding regions of possible pseudogenes. Except for the initial TBLASTN search, which was done using the Celera server, all other analysis steps were automated by investigator-develop-ed programs. (For details of the strategy, see Methods and see Supplementary Methods on the supplementary information page of *Nature Neuroscience* online.)

From this comprehensive data mining effort, we identified nearly the complete mouse olfactory receptor repertoire, which consists of 1,296 genes (including 96 for which only partial sequences were available because the Celera sequence is not yet complete). This constitutes by far the largest gene superfamily in the mammalian genome. About 80% (~1,000) of the OR genes are potentially functional genes, and 20% seem to be pseudogenes. This large number of OR genes is in good agreement with previous predictions based on the number of glomeruli (sensory neuron targets) and from screening genomic phage libraries[5]. The sequences for all receptors obtained through the Celera Assembled and Annotated Mouse Genome are available in GenBank and the Olfactory Receptor Database (ORDB, http://senselab.med.yale.edu/senselab/ordb/).

The Celera Mouse Genome claims >99% coverage, but there are many low-quality sequence regions (long stretches of ambiguous nucleotides), especially in long scaffolds. There may be undiscovered OR sequences in these low-quality sequence regions, as well as in remaining gaps. To determine how much of the mouse OR repertoire we had covered, we compared OR protein sequences accessible in public databases with our database. Over 90% of the mouse ORs in current public databases are also in our database (**Table 1**). Some OR sequences in the public database could, however, be PCR artifacts—'chimera receptors' resulting from recombination among highly related sequences[7]. We also found a few examples of non-OR sequences improperly labeled as OR genes. In ORDB[13], for instance, the genes ORL248, ORL834, ORL844, ORL837 are probably not OR genes. When compared to profile HMMs trained on intact human OR genes, they have large E-values (>30, whereas typical OR genes have E-values $<10^{-10}$). Conservatively, it seems safe to say that our OR database covered more than 90% of the whole mouse OR repertoire.

It should be noted that different mouse strains sometimes have slightly different sequences for the same OR. The Celera mouse genome was assembled from four different strains (129X1/SvJ, 129S1/SvImJ, DBA/2J, and A/J). OR genes in the assembled genome represent the consensus of the strains for which sequence was available for that region. The real OR sequences in each strain might be slightly different, but generally should be >99% identical to the sequence in the current database. All of the sequences in our database were <98% identical with each other, except in a few cases where two very similar genes were unambiguously located at different genomic locations.

### Phylogenetic classification of mouse OR genes

To further classify the OR sequences, we generated a multiple alignment of the 1,296 protein sequences and built a consensus phylogenetic tree from 1,000 bootstrap repetitions. On the basis of the consensus tree, we classified the OR genes into families using a rule whereby all family members must comprise a strong phylogenetic cluster (i.e., a reliable clade, generally possessing >50% bootstrap support) and have more than 40% protein identity. By this definition, mouse OR genes were classified into 228 families containing from 1 to 50 member genes. Because the complete tree of 1,296 OR genes cannot be clearly shown on one page, we built a phylogenetic tree using the consensus sequence for each family (**Fig. 1a**). The OR sequences clearly separate into two broad classes, each with excellent bootstrap support. This is the same Class I and Class II distinction as reported previously for the human OR sequences[8]. Class I receptors resemble the family that was first found in fish[14] and in the frog[15], but had been considered an evolutionary relic in mammals[16]. We developed a classification for the mouse OR families, based on the phylogenetic tree, in which Class I OR families were given family numbers lower than 100 (currently 1–42) and Class II OR families were given family numbers higher than 100 (currently 101–286). If new families are discovered, they can be assigned numbers following the same rule. The number of genes in each family is shown in Fig. 1b.

We propose the following nomenclature system for the mouse OR genes: the prefix 'MOR', followed by the family name (Arabic number 1–42, 101–286, as above), a hyphen (-) and a number

**Table 2a. OR sequences with at least one match for expression data[a].**

| No. of disruptions | 0 | 1 | 2 | >2 |
|---|---|---|---|---|
| Total OR sequences | 904 | 177 | 70 | 145 |
| OR sequences with match | 128 | 27 | 7 | 12 |
| (OR sequences with match)/ total | 14% | 15% | 10% | 8.3% |

[a]From EST database or a cDNA source.

**Table 2b. Number of intact genes and pseudogenes from each group of OR sequences according to length and number of disruptions.**

| Length | All | Full | Full | Full | Partial | Partial | Partial |
|---|---|---|---|---|---|---|---|
| Disruptions | all | 0 | 1 | >2 | 0 | 1 | >2 |
| Total sequences | 1,296 | 875 | 138 | 129 | 29 | 39 | 86 |
| Intact genes | 1036 | 873 | 134 | 0 | 29 | 0 | 0 |
| Pseudogenes | 260 | 2[a] | 4[a] | 129 | 0 | 39 | 86 |

[a]Labeled as pseudogenes because they lack one or more OR signature motifs.

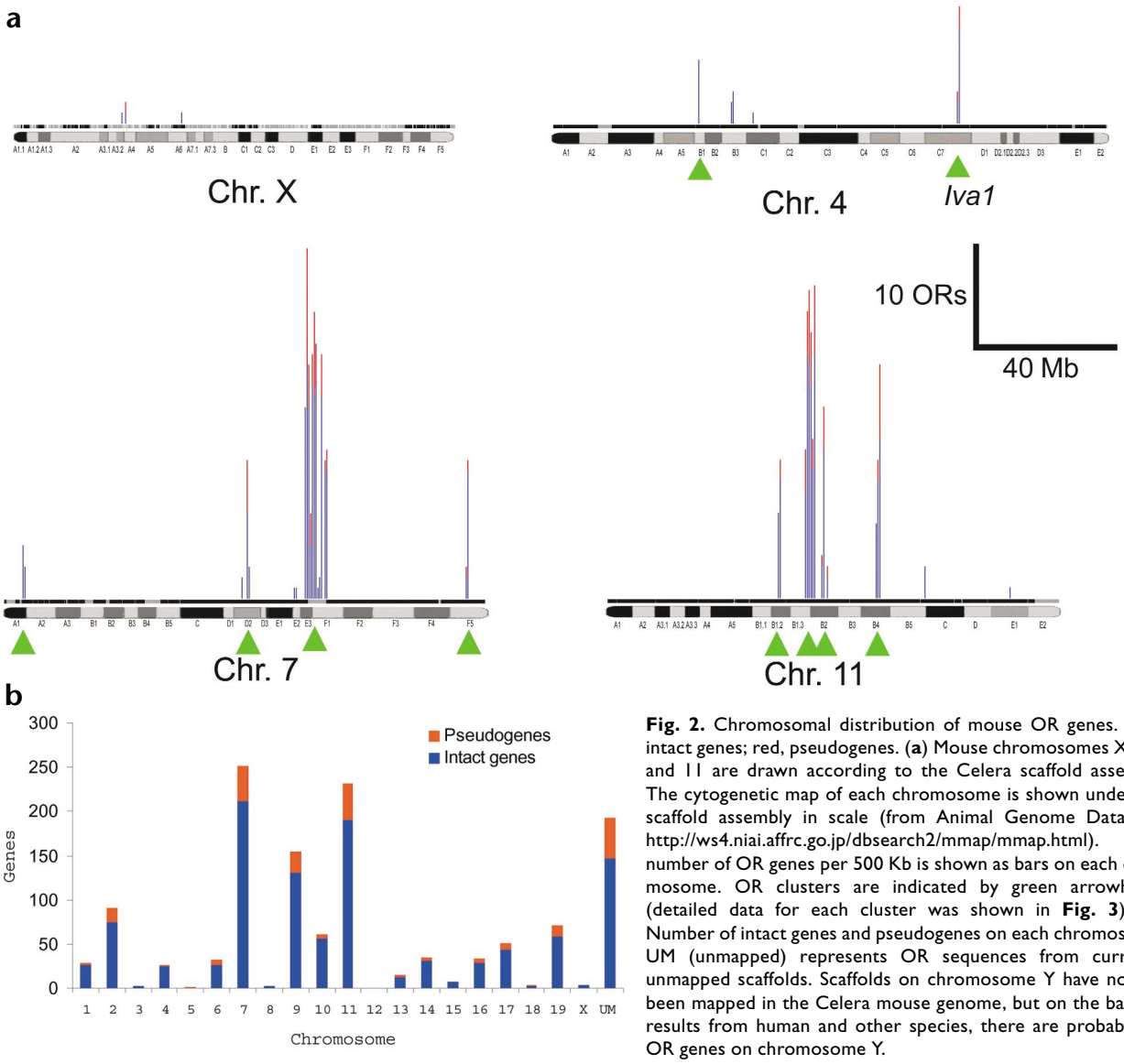


**Fig. 2.** Chromosomal distribution of mouse OR genes. Blue, intact genes; red, pseudogenes. (**a**) Mouse chromosomes X, 4, 7 and 11 are drawn according to the Celera scaffold assembly. The cytogenetic map of each chromosome is shown under the scaffold assembly in scale (from Animal Genome Database, http://ws4.niai.affrc.go.jp/dbsearch2/mmap/mmap.html). The number of OR genes per 500 Kb is shown as bars on each chromosome. OR clusters are indicated by green arrowheads (detailed data for each cluster was shown in **Fig. 3**). (**b**) Number of intact genes and pseudogenes on each chromosome. UM (unmapped) represents OR sequences from currently unmapped scaffolds. Scaffolds on chromosome Y have not yet been mapped in the Celera mouse genome, but on the basis of results from human and other species, there are probably no OR genes on chromosome Y.

representing the individual gene within the family. The letter 'P' at the end of the name denotes a possible pseudogene (see below for discussion of the identification of pseudogenes). For example, MOR1-1 is an intact family 1 OR gene belonging to Class I, because the family number is less than 100; MOR185-9P is a pseudogene in family 185, which is a Class II family.

## Pseudogenes

Of the 1,296 OR genes, we identified a substantial number that had one or more disruptions in the coding region, including insertions, deletions, frame shifts and premature stop codons. We do not believe that all genes with disruptions are pseudogenes, however, as some of the disruptions could be due to errors in the genomic sequences. Indeed, some OR sequences that are known to be functionally expressed had apparent disruptions according to the Celera genome sequences. Aside from possible sequencing errors, there could also be polymorphisms in which one OR gene might contain disruptions in some individuals but not others[6,17,18].

To guide our identification of pseudogenes, we used the available expression data for mouse ORs and compared our OR database to the ORDB[13] (containing 96 mouse OR sequences from cDNA sources) and the mouse EST database (National Center for Biotechnology Information, ftp://ftp.ncbi.nih.gov/blast/db/est_mouse.Z). In total, we found 174 OR sequences with at least one match to cDNA or EST sequences, suggesting functional expression. The percentage of expressed ORs was similar for sequences with no disruptions and with one disruption (14% and 15%, respectively) but little more than half as high for OR sequences with two or more disruptions (8–10%) (**Table 2a**). We therefore estimated that many OR sequences with one disruption could be functional. By contrast, OR sequences with an intact reading frame could be pseudogenes if they lack crucial functional regions. We therefore checked OR sequences with zero or one disruption for the presence of the conserved motifs found in all mammalian ORs[5]. Six lacked one or more of these motifs and were classified as pseudogenes. In summary, we labeled as pseudogenes full-length OR sequences with two or more disruptions, partial-length OR


**Fig. 3.** OR gene clusters, with detailed distribution of OR genes in each of the 27 clusters. OR families (1–42 for Class I and 101–286 for Class II) are color coded as in Fig. 1. X, intact genes; +, pseudogenes. Similar OR genes tend to be located together, forming color patches in the figure. All the Class I OR genes (labeled in gray-scale shades) are in cluster 7-3. Most families are located in a restricted area in one cluster.



sequences with one or more disruptions, and OR sequences with less than one disruption but missing conserved motifs (**Table 2b**). Using these criteria, we classified 260 of the 1,296 OR sequences in mouse (20%) as pseudogenes. If a more conservative approach were taken and only full-length genes with no disruptions were considered functional genes, there would be only 873 functional genes and 423 pseudogenes (33%).

## Genomic distribution

OR genes were distributed mainly in clusters on all mouse chromosomes except 12 and Y. Of the 1,296 OR genes, 1,103 were mapped to 18 chromosomes, and the rest were in currently unmapped sequence regions (**Fig. 2**). Chromosome 7 housed the largest number of OR genes (252), with chromosomes 11 (190) and 9 (131) second and third, respectively. In contrast, chromosomes 3 and 8 had only 2 OR genes apiece, and chromosome X only 4 (**Fig. 2b**).

Gene clusters were determined using the same definition as in the human OR genome: clusters contained more than five genes, none separated by >1 Mb[8]. We identified 27 mouse OR clusters—including two on currently unmapped scaffolds that could be part of existing clusters—containing a total of 1,130 OR genes (**Fig. 3**). The frequency of OR genes in these clusters was high, ranging from 18 to 66 Kb per OR gene (average, 29 Kb per OR gene). OR clusters were named according to their chromosome number (1–19, and UM for 'unmapped') and the index number of the cluster on the chromosome. OR genes from the same family tended to be located near each other, forming 'subclusters'. Some OR genes were not located in clusters; however, many of these were in 'miniclusters' consisting of fewer than six genes together (as in cytogenetic band B3 of chromosome 4; **Fig. 2a**). There were few isolated, single OR genes; one was located in the middle of chromosome X.

In spite of the density of OR genes, non-OR genes were regularly found within the OR clusters. Only five small OR clusters (1-1,4-1, 4-2, 7-1, UM-1) were completely free of non-OR genes; all other clusters had some non-OR genes distributed within them. Genes encoding viral coat proteins (Gag, Env and Pol polyproteins) were often found in OR clusters. Notably, the genes encoding retrovirus-related Gag or Gag-related proteins could be found in 15 of the 27 OR clusters, present in 1–8 copies. The density of such genes was twice as high in OR clusters as in the rest of the genome. The presence of viral coat proteins in OR clusters suggests a possible viral-based mechanism of gene duplication and relocation. OR genes, like those of many mammalian GPCRs, are generally intronless, and one theory attributes this to retroviral-mediated duplication of the family[19,20].

To determine if phylogenetically related OR genes are located close to one another, we identified OR pairs that were at least 60% identical at the protein level and investigated their relative chromosomal locations. Of these, 1,176 OR genes had another OR gene that was at least 60% identical, but because 239 of these pairs had at least one OR gene not mapped to a chromosome location, only 937 were analyzed. In 918 (98.0%) of these cases, the two OR genes were located on the same chromosome, within a median distance of 44 Kb. In 55.3% (518/937), the two OR genes were either adjacent to each other or separated by only one other OR gene. This close proximity of highly related OR genes suggests local duplication as another mechanism of OR family expansion.

## Candidate OR cluster for an anosmia to isovaleric acid

Several naturally occurring specific anosmias (inability to detect a particular odor) have been reported in human and mouse[21].

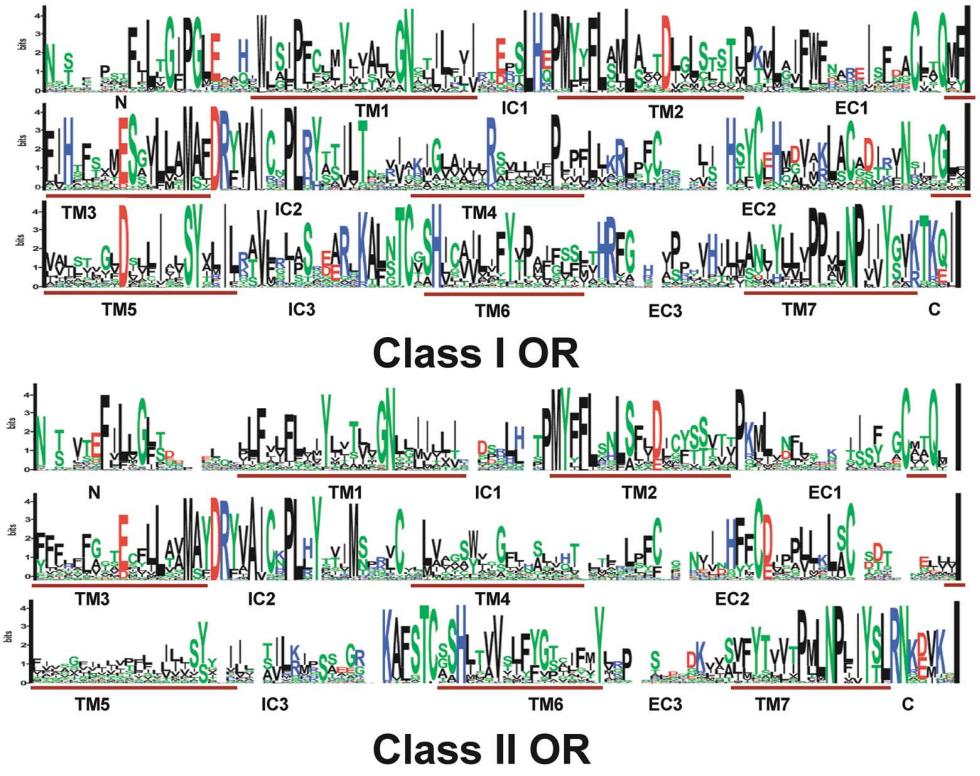


**Fig. 4.** Sequence logos for the open reading frames of Class I and Class II ORs. The very terminal sequences are removed to avoid length heterogeneity; no significant sequence conservation was found in these regions. The height of each amino acid is proportional to its frequency of occurrence. Locations of predicted transmembrane regions (TM1–7), intracellular loops (IC1–3) and extracellular loops (EC1–3) are shown. The large number of genes tends to reduce the conservation of any given residue. The characteristic OR sequences can still easily be seen.

C57BL/6J and C57BL/10J mice have a specific anosmia (or more precisely, a narrowly limited hyposmia) to isovaleric acid, and this defect seems to have a peripheral source[22,23]. This anosmia is recessive, and two loci responsible for the defect have been mapped, *Iva1* on chromosome 4 and *Iva2* on chromosome 6 (ref. 21). On the basis of adjacent molecular markers, we located *Iva1* at 110.46M–112.23M on the reference axis of chromosome 4 (**Fig. 2a**), a location that matches very well with OR cluster 4-2 (111.96M–112.29M on chromosome 4). The 14 OR genes in this cluster were from two closely related families, 258 and 259, that form a clade with 97% bootstrap support value. Two other unmapped OR genes also belong to these two families. Because highly related sequences were usually located near one another, we suspected that the two unmapped OR genes probably were also part of cluster 4-2. No other OR genes from any other position fell into these two families. Thus we considered the 16 genes in families 258 and 259 as *Iva1* OR genes. The other locus, *Iva2*, was mapped to 136.1M–140.9M of chromosome 6, but no OR sequences were found in this region. There is a sequence gap in the Celera mouse genome in this region, so there may be OR genes in this gap. Considering, however, *Iva2* that is only weakly correlated with the anosmia and that one of the markers (*D6MIT201*) used to locate *Iva2* actually maps to the *Iva1* locus in the Celera mouse genome, it seems probable that *Iva2* is not the true locus for isovaleric acid anosmia. The most likely cause of the anosmia in C57BL/6J and C57BL/10J mice seems to be the loss of OR proteins in cluster 4-2. Because the strains used in constructing the Celera mouse genome database were osmic, it would be interesting to sequence the *Iva1* locus of the anosmic strains for evidence of the genetic lesion underlying the loss of these OR proteins.

## Global view of sequences

We generated 'logo' views of sequences of the Class I and Class II ORs to facilitate visual sequence comparisons, showing predicted transmembrane (TM), intracellular (IC) and extracellular (EC) regions (**Fig. 4**). Both classes showed the characteristic olfactory receptor–specific regions (e.g., MAYDRYVAIC in TM3-IC2, FSTCSSH in IC3-TM6 and the three conserved cysteines in EC2). Some features were quite distinct between the two classes, however. For example, the MAYDRY motif was more often MAFDRY in Class I ORs; there were three conserved prolines in TM7 of Class I ORs, but only two prolines in Class II ORs, and a highly conserved region starting from the middle of IC2 to the middle of TM4 (M....C..Lv...sW) was present in almost all Class II ORs but absent in Class I ORs.

Transmembrane domains 4 and 5, and to lesser extent TM3, are the most variable regions of ORs[24]. This can be seen in the logo view, where fewer conserved residues were seen in these regions. We expect that further analyses of conserved and variable regions of the mammalian ORs will identify key regions that may be instructive for functional studies of ORs in particular and GPCRs in general.

## Comparison with other mammalian species

Cross-species comparison of ORs is often subject to the problem that the closest match is not in the available database (typically the ORDB) and thus a false distant ortholog is retrieved instead. As our mouse OR database covers almost all the mouse ORs in the genome, we expected that cross-species comparisons would be essentially free of this problem.

The species most closely related to mouse for which numerous sequences are available in the ORDB is the rat. Isolated examples

**Table 3. Cross-species matches with mouse OR database.**

| Species | >40% identity | >60% identity | >80% identity | >90% identity | >95% identity |
|---|---|---|---|---|---|
| Human | 100% (347/347) | 93% (323/347) | 58% (200/347) | 5% (18/347) | 0.3% (1/347) |
| Rat | 100% (65/65) | 98% (64/65) | 91% (59/65) | 55% (36/65) | 17% (11/65) |

Intact human (347) and rat ORs (65) were compared with mouse OR database, and the protein identity of the closest match of each OR was used for calculation.

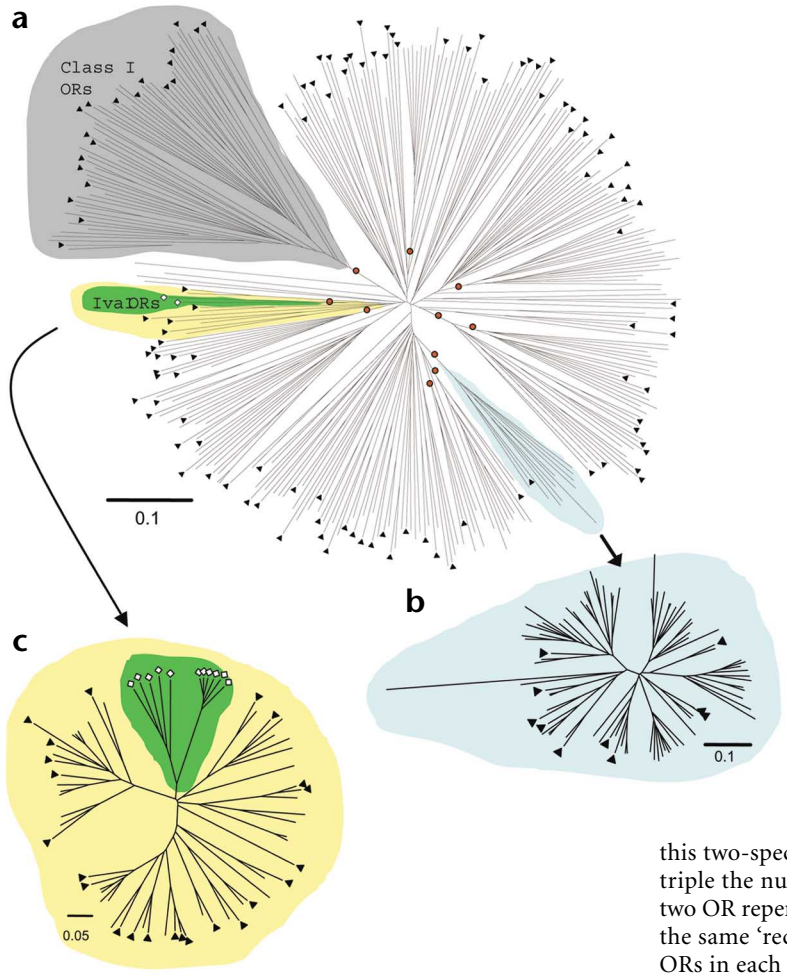


**Fig. 5.** Unrooted phylogenetic tree of human and mouse ORs. (**a**) Unrooted phylogenetic tree of the consensus protein sequences of mouse and human OR gene families. Human OR families are indicated by filled triangles. A few groups (clades) with high (>90%) bootstrap value are labeled by red dots. The Class I OR families, shaded in gray, form a group with more than 90% bootstrap value. A group with high bootstrap value and including mostly mouse OR families is shaded with light blue. The *Iva1* ORs (mouse OR families 258 and 259) are indicated by open diamonds and shaded with green; families close to the *Iva1* OR families are shaded with yellow. (**b**) Unrooted phylogenetic tree of the intact full-length ORs belonging to families in the group shaded in light blue in (**a**), which contains 11 mouse OR families and only 1 human family. Although there are only 10 intact human ORs in this group, they do not form one tight subgroup but intermingle with the 76 intact full-length mouse ORs covering more than half of the subgroups. (**c**) Unrooted phylogenetic tree of the intact full-length ORs of the *Iva1* ORs (mouse OR families 258 and 259), shaded in yellow in (**a**), and families close to them. *Iva1* ORs form two families, which compose a group with high bootstrap value. None of the human ORs can be found in these two families, although they intermingle with other mouse families.



had indicated that the two species might have very similar OR repertoires (e.g., rat I7 and mouse I7 receptors are 95% identical[25]). We found that the two species did have similar OR repertoires: 90% of the 65 rat ORs have a mouse ortholog with more than 80% identity. However, 45% of the rat ORs did not have a mouse ortholog with >90% identity, indicating that there are also considerable differences between the repertoires (**Table 3**). Precisely where the sequences diverge may provide useful hints to functional differences between receptors in the two species.

Being more distant from mouse, humans have substantially different OR repertoires. Although all intact human ORs had mouse orthologs of >40% identity, only 59% had an ortholog with >80% identity, and only 4% had one with >90% identity (**Table 3**). One OR gene, *OR93*, has been identified in a few species of apes; it has an ortholog with 96% identity in human, *OR5G1P* (official Human Genome Organisation (HUGO) designation), but the closest match in mouse (*MOR175-1*, or *Olfr4-3* according to Mouse Genome Informatics (MGI)/Jackson Labs designation) had only 80% identity. A few dog ORs seem to fall somewhere in between human and mouse, with about 80% identity to each species.

Because the complete human OR repertoire is available, a comprehensive phylogenetic analysis of the human and mouse OR subgenomes is possible. Some 347 human intact ORs have been classified into 119 families[9], and from this we produced 119 consensus sequences for each of the human OR families. We used these to build a phylogenetic tree of consensus protein sequences of all human and mouse OR families (**Fig. 5a**). In this two-species tree, although the mouse genome had nearly triple the number of intact ORs, the overall structures of the two OR repertoires were similar, and they covered more or less the same 'receptor space'. Although mice generally had more ORs in each broad branch, humans did not seem to have lost any of the major branches. Even in some groups where mouse families predominated, intact human ORs in those groups intermingled with the mouse ORs and covered a relatively broad space (**Fig. 5b**). Thus, the human olfactory system has probably retained the ability to recognize a broad, if perhaps less discriminating, spectrum of chemicals while using one-third the number of ORs as in mouse.

There were a few groups of ORs present only in mouse, such as the two families in the *Iva1* locus (**Fig. 5c**). These exclusive mouse groups were uncommon and were typically small. Given that olfactory coding is probably primarily combinatorial, using numerous ORs to recognize an odor compound, the ORs not present in humans probably alter sensitivity or discrimination but not the range of detectable compounds.

## Discussion

We have carried out a comprehensive search of the mouse genome recently released by Celera for genes encoding olfactory receptors (ORs). From the earliest cloning data, the size of this gene family was estimated to be on the order of 1,000 genes, by far the largest single gene family encoding neural genes. In fact, our data show an even larger family of receptors consisting of 1,296 genes distributed in clusters of varying sizes throughout the 18 of the 20 mouse chromosomes. These genes also constitute the largest expansion of the family of GPCRs, an already large and important superfamily of membrane receptors with diverse ligands and functions. Although the Celera mouse genome is in its first iteration and there are gaps and

regions of low-quality sequence, by comparing our database with previously known OR genes, we conservatively estimate that we have identified >90% of the OR genes.

Previously only a few full-length OR sequences had been available in any mammal. In conjunction with the human OR gene family[8,9], we now have a very large number of full-length gene sequences, from which several new insights have emerged.

### Classification based on phylogenetic relations

The enormous size of the OR gene family makes its classification into manageable units especially crucial. By multiple alignments of full-length sequences, we defined 228 families with good bootstrap support. Families contained from 1 to 50 genes, all of which were at least 40% identical at the protein level and which formed a reliable clade. We chose a nomenclature system that is easy to use and makes no assumptions regarding receptor function (such as ligand preferences). Genes are named by family number (1–99 for Class I families, above 101 for Class II) and assigned an individual number within the family.

Two nomenclature systems exist for human ORs. Lancet and colleagues proposed naming OR genes according to family (>40% amino acid identity) and subfamily (>60%) assignments[26], whereas Zozulya and colleagues introduced an alternative scheme based mainly on phylogenetic analysis but also accounting for genomic location[9]. Our proposed nomenclature is more like the second system, except that we did not include genomic location. We did this because some families contained member genes from multiple chromosomes and because there were nearly 200 OR genes not currently mapped to genomic locations. We examined the possibility of using the Lancet nomenclature system for mouse OR genes, but found that some of the resulting families and subfamilies had very low bootstrap values, making the classification phylogenetically unreliable.

### A large number of pseudogenes in mouse OR sequences

Among human OR sequences, a marked proportion—nearly 65%—are classified as pseudogenes[8]. It has been thought that this situation might be unique to the human repertoire and that in other mammals the frequency of pseudogenes would be much lower, presumably as a result of greater selective pressure associated with olfactory abilities in mammals other than humans.

It is not always easy to determine whether a gene is functional. In the mouse repertoire, we recognized a gene as a pseudogene if it met one of the three criteria: two or more disruptions in a full-length gene, fewer than two disruptions and the absence of any of the highly conserved motifs found in OR genes, or partial sequence with one or more disruption. According to these criteria, about 20% of the OR sequences in mouse (260 genes) were identified as pseudogenes. Although notably lower than in humans, this is a higher value than expected. Less than 10 pseudogenes have been previously reported in mouse[10,27], a fact attributable to two factors. First, most OR sequences have been obtained from cDNA, which would miss untranscribed genes, and second, the common practice of using degenerate PCR to obtain OR partial sequences would allow disruptions outside the PCR-amplified region to be missed.

The large number of pseudogenes suggests that the OR superfamily undergoes rapid evolution, with new genes continuously being generated by duplication and mutation. Indeed, as a result of multiple recent mutation events, the actual number of pseudogenes could be even higher, as heavily disrupted ORFs may not be recognizable as OR genes by a standard TBLASTN search.

We expect that, for mouse ORs, something less than the ~1,000 apparently intact genes are functional in any individual. From limited data, it seems that, on average, the cells expressing a given receptor target two glomeruli in the mouse olfactory bulb. Counting glomerular targets (~1,800) and dividing by 2 leads to a rough estimate of 900 functional genes[5]. The number of glomeruli in the human olfactory bulb is not well established, nor is it known whether the same targeting ratio applies in the human system; therefore, it is impossible to make a similar estimate in humans. In the human genome, however, more than 50 genes with only one disruption are currently labeled as pseudogenes; by the reasoning that we have applied in mouse, many of these could be intact genes.

### OR gene distribution and comparison with known loci

In addition to the phylogenetic families, we found that OR genes were distributed on all mouse chromosomes except 12 and Y. The distribution was not uniform, with more than half of the genes contained in a few large, compact clusters on chromosomes 7, 11 and 9. In general the clusters are densely packed with OR genes, although non-OR genes do occur within clusters. In addition, phylogenetically related OR genes were usually in close proximity to one another, suggesting that the expansion of the OR family is due not only to large-scale duplications of clusters but also to local duplication within clusters.

A few mouse OR loci have been genetically mapped to chromosomal locations[28–34]. We matched these loci to our database using either their OR sequences or molecular marker sequences. The same chromosome locations were found for all of the known mouse OR loci except *olfr4* (**Fig. 3**). Although only one or a few OR genes were previously known for each locus, we mapped them to OR clusters with dozens of OR genes. Not surprisingly, quite often several known loci mapped to different locations within the same cluster.

Two OR clusters have recently been studied in detail. In one case, 18 mouse OR genes were found in *olfr17* (ref. 12), which matches to one subcluster (87.5M–87.9M) of 25 OR genes in cluster 7–3 in our data. Some 42 OR genes have been identified in *olfr7*, and the total number of OR genes in this cluster was estimated to be around 100 (ref. 10). This group matches cluster 9–2 in our data, which in fact has 113 OR genes.

*Olfr4* was previously mapped to chromosome 2 and has three known genes. However, we mapped one gene to cluster 11–2 and two genes to an unmapped scaffold, UM–2, that was also probably on chromosome 11. Resolving this discrepancy in location will require additional data.

One locus, *Iva1*, has been identified with a specific anosmia to isovaleric acid. We mapped two families of receptor genes containing a total of 16 OR genes to that precise location, indicating that some or all *Iva1* ORs bind isovaleric acid with high affinity. Mice from anosmic strains respond to isovaleric acid at a higher threshold than mice from osmic strains[23], indicating that the defect might involve a loss of high-affinity receptors. Of the 16 *Iva1* OR genes, six intact genes were in family 258, sharing 50–80% protein identity, and seven intact genes plus three pseudogenes were in family 259, sharing 80–95% protein identity. Genes from the two families shared 35–50% protein identity, indicating that they may have different ligand binding profiles. Functional studies of these two families could reveal which family, or which ORs, are responsible for recognizing isovaleric acid at high affinity.

When the *Iva1* ORs were compared with human ORs, no intact human OR fell into the same clade as the *Iva1* ORs (**Fig. 5c**). At least from the available data, no orthologs of *Iva1* ORs are present in human, which suggests that humans lack the

high-affinity receptors for isovaleric acid. In animal behavior assays in which isovaleric acid was diluted in buffered solution adjusted to its pK, anosmic mice could detect isovaleric acid only at concentrations higher than $10^{-5}$ M, whereas osmic strains were sensitive to isovaleric acid at concentrations as low as $10^{-7}$ M[21]. According to Flavor-Base Pro (Leffingwell & Associates, Canton, Georgia), the human flavor threshold for isovaleric acid in water is 120–700 ppb, which is equivalent to $1.2\times10^{-6}$–$6.9\times10^{-6}$ M, a number between the thresholds of anosmic and osmic mice. Strictly controlled experiments under the same conditions as those used in animal assays would be required to determine the precise human threshold for the detection of isovaleric acid.

### Class I ORs

Class I ORs, first identified in fish[14], separate clearly in the phylogenetic tree from the classical, mammalian-specific Class II ORs. There were 147 Class I OR genes in the mouse OR subgenome, 120 of them potentially functional (**Fig. 1**). All of the Class I OR genes were located in a single large cluster on chromosome 7 (cluster 7-3; **Fig. 3**). Class I ORs were previously thought to be evolutionary relics in mammals[16]; however, there are a relatively large number of intact Class I OR genes in the human genome[8], and we found an even larger number in the mouse genome. This confirms that Class I OR genes are prevalent in the mammalian genome and indicates that they may be centrally involved in mammalian olfaction. Expression data is available for some of the mouse and rat Class I ORs, and in all cases they are expressed in the most dorsal zone of the olfactory epithelium[33,35–37]. In contrast, Class II receptors have been found in all four zones. It should be noted that this distinction is based on limited data for only a few receptors, and establishing a differential expression pattern for these two classes of receptors would require considerably more data.

Notably, 11 of 14 ORs recently identified as having aliphatic odor ligands[36] belong to Class I. In that study, odorant compounds were applied to dissociated olfactory neurons plated on coverslips, and responses were monitored by calcium imaging. The large proportion of Class I ORs found in this study is unusual (11/14 compared to their 12% occurrence in the whole OR repertoire), which indicates that the experimental design or the compounds used in the study might have favored the activation of olfactory neurons expressing Class I ORs. The first possibility can be eliminated, as similar experiments using different odorant compounds did not isolate a large proportion of Class I ORs[38,39]. It therefore seems likely that the aliphatic compounds used in this study—all of them acids and alcohols, which are relatively hydrophilic compared to many other odorant compounds—were mostly Class I OR–specific ligands.

Class I ORs are related to fish ORs, which are expected to bind water-soluble compounds. In frog, Class I ORs are activated by water-soluble odorants, whereas Class II ORs are activated by volatile compounds[40]. In mammals, however, water-soluble compounds generally are not strong odorants, and Class I ORs are considerably divergent from those of fish or frog. We hypothesize that Class I ORs in mammals might have evolved to recognize volatile compounds, although they are still more sensitive to relatively hydrophilic compounds, whereas Class II ORs might favor more hydrophobic compounds.

In summary, we have identified some ~1,300 OR sequences from the nearly complete Celera mouse genome. There are about 1,000 functional OR genes, roughly correlating with the number of glomerular targets in the olfactory bulb. The high percentage of pseudogenes (~20%) was unexpected and suggests that OR genes are one of the fastest-evolving gene families. As previously found in human, a large number of 'fish-like' Class I OR genes exist in the mouse genome, confirming their functional role in mammalian olfaction. Using the genomic distribution data, we were able to map a specific anosmia locus to an OR cluster, the first time a specific anosmia has been directly associated with a group of OR genes. By comparing the mouse and human OR genes at the whole-genome level, it seems that humans have retained the ability to detect a wide range of odorants with one-third the number of receptors by keeping fewer ORs in each family. The size of the mouse OR repertoire effectively increases the number of available GPCR sequences by at least twofold, adding considerably to the database of functional GPCRs. These data should prove useful for efforts to develop an understanding not only of how animals sense their chemical environment but also of the mechanisms by which GPCRs recognize a wide range of ligands.

## METHODS

**Data mining.** An exhaustive TBLASTN search incorporating profile HMM search was used to obtain all the possible OR sequences from the Celera mouse genome. Conceptual translation was used to recover the original ORFs for possible pseudogenes. Duplicates were removed, and the resulting OR genes were mapped to genomic locations according to the mapping data of the scaffolds by Celera. (For detailed description, see Supplementary Methods and **Supplementary Figs. 1** and **2** on the supplementary information page of *Nature Neuroscience* online.)

**Matches with known OR genes.** We collected 122 mouse OR genes from the ORDB and 362 mouse OR genes from GenBank using 'olfactory receptors' or 'odorant receptors' as a keyword and searching mouse genes (all databases as of 6/25/2001). The protein sequences encoded by the OR genes from the public database were matched with our mouse OR database using FASTA3. For each OR gene from the public database, the best hit was chosen and the percentage of protein identity was used for further analysis. Similarly, human and rat OR genes were also matched with our mouse OR gene database.

**Matches with OR sequences from cDNA sources and the mouse EST database.** OR sequences labeled as originating from cDNA source material in the ORDB were selected, and each was searched against our mouse OR database. Hits with >95% identity were considered matches. The mouse EST database was downloaded from the NCBI server, and a BLASTN search was done using every mouse OR DNA sequence as a query against the EST database. Hits with E-values $<1^{-100}$ were considered matches.

**Phylogenetic analysis of the OR sequences.** The protein sequences encoded by the 1,296 mouse OR genes were aligned using ClustalX 1.81. The resulting multiple alignment were used as input to PAUP*4.0 beta (Sinauer Associates, Sunderland, Massachusetts) and the majority-rule consensus neighbor-joining (NJ) tree from 1,000 bootstraps was obtained from PAUP*, requiring 48 h of running time on a 1-GHz Pentium III PC. OR gene families were determined from the tree as the largest clades that fulfilled two criteria: the clade had >50% bootstrap support, and all members within the clade had at least 40% protein identity.

To show a simplified tree, we built another tree with the consensus sequences of each family. Only intact, full-length OR genes from each family were used to generate the consensus sequence. The sequences from each family were aligned using ClustalW, a profile HMM was built upon the alignment, and the consensus sequence was generated from the profile HMM using the HMMER package. The OR genes from families with only a single gene were used directly. All the consensus sequences were aligned, and an NJ tree with 1,000 bootstraps was built using ClustalX. The tree was rooted using human and mouse melanocyte-stimulating hormone receptors (MSH-R), one of the GPCRs most closely related to ORs. The tree was plotted using Tree Explorer (http://evolgen.biol.metrou.ac.jp/TE/TE_man.html).

The same method was used to obtain human consensus sequences and to build a combined tree of all human and mouse OR families. The tree

was unrooted and plotted using Tree Explorer. Selected OR genes from human and mouse (**Fig. 5b** and **c**) were aligned and NJ trees with 1,000 bootstraps were built using ClustalX.

**Sequence logos.** Clustal X 1.81 was used for multiple alignments of full-length intact Class I and Class II ORs. The alignments were manually edited. Gap positions present in >98% of the sequences were deleted. Sequence logos were generated using a web-based program developed by J. Gorodkin (www.cbs.dtu.dk/gorodkin/appl/plogo/html). The secondary structure prediction was based on the PredictProtein Server results on consensus sequences generated by the HMMER package.

**GenBank accession numbers.** OR gene sequences from the Celera database: AY072961-AY074256. Ape *OR93*: AAC63969–63971.

*Note: Supplementary information is available on the Nature Neuroscience web site (http://neuroscience.nature.com/web_specials).*

## ACKNOWLEDGMENTS

*We wish to thank P. Feinstein, P. Mombaerts and the members of the Firestein lab for critical comments on the manuscript. This work was supported by grants from US NIDCD and HFSP. This data was generated through use of the Celera Discovery System and Celera's associated databases made possible in part by the AMDeC Foundation, Inc.*

RECEIVED 21 SEPTEMBER; ACCEPTED 23 DECEMBER 2001

1. Gibson, A. D. & Garbers, D. L. Guanylyl cyclases as a family of putative odorant receptors. *Annu. Rev. Neurosci.* **23**, 417–439 (2000).
2. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).
3. Mombaerts, P. Seven-transmembrane proteins as odorant and chemosensory receptors. *Science* **286**, 707–711 (1999).
4. Buck, L. B. Information coding in the vertebrate olfactory system. *Annu. Rev. Neurosci.* **19**, 517–544 (1996).
5. Mombaerts, P. Molecular biology of odorant receptors in vertebrates. *Annu. Rev. Neurosci.* **22**, 487–509 (1999).
6. Glusman, G. *et al.* Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. *Genomics* **63**, 227–245 (2000).
7. 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* **37**, 147–160 (1996).
8. Glusman, G., Yanai, I., Rubin, I. & Lancet, D. The complete human olfactory subgenome. *Genome Res.* **11**, 685–702 (2001).
9. Zozulya, S., Echeverri, F. & Nguyen, T. The human olfactory receptor repertoire. *Genome Biol.* **2**, 0018.0011–00180012 (2001).
10. Xie, S. Y., Feinstein, P. & Mombaerts, P. Characterization of a cluster comprising approximately 100 odorant receptor genes in mouse. *Mamm. Genome* **11**, 1070–1078 (2000).
11. Hoppe, R., Weimer, M., Beck, A., Breer, H. & Strotmann, J. Sequence analyses of the olfactory receptor gene cluster mOR37 on mouse chromosome 4. *Genomics* **66**, 284–295 (2000).
12. Lane, R. P. *et al.* Genomic analysis of orthologous mouse and human olfactory receptor loci. *Proc. Natl. Acad. Sci. USA* **98**, 7390–7395 (2001).
13. Skoufos, E. *et al.* Olfactory Receptor Database: a database of the largest eukaryotic gene family. *Nucleic Acids Res.* **27**, 343–345 (1999).
14. Ngai, J., Dowling, M. M., Buck, L., Axel, R. & Chess, A. The family of genes encoding odorant receptors in the channel catfish. *Cell* **72**, 657–666 (1993).
15. Freitag, J., Krieger, J., Strotmann, J. & Breer, H. Two classes of olfactory receptors in *Xenopus laevis*. *Neuron* **15**, 1383–1392 (1995).
16. Freitag, J., Ludwig, G., Andreini, I., Rossler, P. & Breer, H. Olfactory receptors in aquatic and terrestrial vertebrates. *J. Comp. Physiol. [A]* **183**, 635–650 (1998).
17. Ehlers, A. *et al.* MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes. *Genome Res.* **10**, 1968–1978 (2000).
18. Younger, R. M. *et al.* Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res.* **11**, 519–530 (2001).
19. Brosius, J. Many G-protein-coupled receptors are encoded by retrogenes. *Trends. Genet.* **15**, 304–305 (1999).
20. Gentles, A. J. & Karlin, S. Why are human G-protein-coupled receptors predominantly intronless? *Trends. Genet.* **15**, 47–49 (1999).
21. Griff, I. C. & Reed, R. R. The genetic basis for specific anosmia to isovaleric acid in the mouse. *Cell* **83**, 407–414 (1995).
22. Wysocki, C. J., Whitney, G. & Tucker, D. Specific anosmia in the laboratory mouse. *Behav. Genet.* **7**, 171–188 (1977).
23. Wang, H. W., Wysocki, C. J. & Gold, G. H. Induction of olfactory receptor sensitivity in mice. *Science* **260**, 998–1000 (1993).
24. Pilpel, Y. & Lancet, D. The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8**, 969–977 (1999).
25. Krautwurst, D., Yau, K. W. & Reed, R. R. Identification of ligands for olfactory receptors by functional expression of a receptor library. *Cell* **95**, 917–926 (1998).
26. Glusman, G. *et al.* The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mamm. Genome* **11**, 1016–1023 (2000).
27. Lapidot, M. *et al.* Mouse–human orthology relationships in an olfactory receptor gene cluster. *Genomics* **71**, 296–306 (2001).
28. Copeland, N. G. *et al.* A genetic linkage map of the mouse: current applications and future prospects. *Science* **262**, 57–66 (1993).
29. Sullivan, S. L., Adamson, M. C., Ressler, K. J., Kozak, C. A. & Buck, L. B. The chromosomal distribution of mouse odorant receptor genes. *Proc. Natl. Acad. Sci. USA* **93**, 884–888 (1996).
30. Carver, E. A., Issel-Tarver, L., Rine, J., Olsen, A. S. & Stubbs, L. Location of mouse and human genes corresponding to conserved canine olfactory receptor gene subfamilies. *Mamm. Genome* **9**, 349–354 (1998).
31. Strotmann, J. *et al.* Small subfamily of olfactory receptor genes: structural features, expression pattern and genomic organization. *Gene* **236**, 281–291 (1999).
32. Asai, H. *et al.* Genomic structure and transcription of a murine odorant receptor gene: differential initiation of transcription in the olfactory and testicular cells. *Biochem. Biophys. Res. Commun.* **221**, 240–247 (1996).
33. Bulger, M. *et al.* Conservation of sequence and structure flanking the mouse and human beta-globin loci: the β-globin genes are embedded within an array of odorant receptor genes. *Proc. Natl. Acad. Sci. USA* **96**, 5129–5134 (1999).
34. Zheng, C., Feinstein, P., Bozza, T., Rodriguez, I. & Mombaerts, P. Peripheral olfactory projections are differentially affected in mice deficient in a cyclic nucleotide–gated channel subunit. *Neuron* **26**, 81–91 (2000).
35. Conzelmann, S. *et al.* A novel brain receptor is expressed in a distinct population of olfactory sensory neurons. *Eur. J. Neurosci.* **12**, 3926–3934 (2000).
36. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96**, 713–723 (1999).
37. Raming, K., Konzelmann, S. & Breer, H. Identification of a novel G-protein coupled receptor expressed in distinct brain regions and a defined olfactory zone. *Receptors Channels* **6**, 141–151 (1998).
38. Touhara, K. *et al.* Functional identification and reconstitution of an odorant receptor in single olfactory neurons. *Proc. Natl. Acad. Sci. USA* **96**, 4040–4045 (1999).
39. Kajiya, K. *et al.* Molecular bases of odor discrimination: reconstitution of olfactory receptors that recognize overlapping sets of odorants. *J. Neurosci.* **21**, 6018–6025 (2001).
40. Mezler, M., Fleischer, J. & Breer, H. Characteristic features and ligand specificity of the two olfactory receptor classes from *Xenopus laevis*. *J. Exp. Biol.* **204**, 2987–2997 (2001).