Chapter 3

Identification of Olfactory Receptor Genes from Mammalian Genome Sequences

Yoshihito Niimura

Abstract

Olfaction is essential for the survival of mammals. Diverse odorant molecules in the environment are detected by olfactory receptors (ORs) expressed in the olfactory epithelium of the nasal cavity. In general, mammalian genomes harbor ~1,000 OR genes, which form the largest multigene family in mammals. The recent advances in genome sequencing technology have enabled us to computationally identify nearly complete repertoires of OR genes from various organisms. Such studies have revealed that the numbers of OR genes are highly variable among organisms depending on their living environments.

Because OR genes are intronless, it is possible to find all OR genes by conducting homology searches against the genome sequences using known OR genes as queries. However, some caution is necessary during the process of extracting intact coding sequences of OR genes and distinguishing among OR and non-OR genes. Presented here is a description of bioinformatics methods to identify the entire OR gene repertoires from mammalian genome sequences.

Key words Olfactory receptor, Multigene family, Bioinformatics, Mammalian genome, G-protein coupled receptor

1 Introduction

Diverse odor molecules in the environment are detected by olfactory receptors (ORs) that are expressed in the olfactory epithelium of the nasal cavity. To deal with enormous diversity of odorants, the mammalian genomes harbor OR genes with varied sequences. The genomes of many mammalian species contain ~1,000 or more OR genes, which form the largest multigene family in mammals (*for reviews*, see (1, 2)). OR genes were first identified from rats by Linda Buck and Richard Axel in 1991 (3), a discovery that led to them being awarded the Nobel Prize in 2004.

An OR is a G-protein coupled receptor (GPCR) with seven α-helical transmembrane (TM) regions. ORs are, on average, approximately, 310 amino acids long. OR genes belong to rhodopsin-like

Chiquito J. Crasto (ed.), *Olfactory Receptors: Methods and Protocols*, Methods in Molecular Biology, vol. 1003,
DOI 10.1007/978-1-62703-377-0_3,

GPCR superfamilies, which include opsins for detecting light and receptors for various ligands such as neurotransmitters, peptide hormones, chemokines, lipids, and nucleotides, and others (4). OR genes do not have any introns in their coding regions. Introns are often present in the 5′ untranslated regions of an OR gene; however, different mRNA isoforms generated by alternative splicing of noncoding exons result in the same protein (5).

OR genes are present in all vertebrate species. Fish have much smaller repertoires of OR genes (~100) than mammals. However, although OR gene repertoires in fish are small, their OR genes are more diverse in sequence than those of mammals (6, 7). Mammalian, reptilian, and avian OR genes can be clearly classified into two groups, Class I and Class II (8), on the basis of their amino acid sequence similarities. On the other hand, based on sequence similarities, amphibians and fish have additional groups of OR genes (6, 7).

Amphioxus, the most basal chordate species, also possesses vertebrate-like OR genes (7). Therefore, the origin of OR genes can be traced back to the common ancestor of all chordate species. Insects and nematodes also have OR genes. However, those in chordates, insects, and nematodes do not show sequence similarities, suggesting multiple origins of OR genes during animal evolution (1, 2, 9).

Thanks to the recent advances in genome sequencing technologies, whole genome sequences of various organisms have become available. We have established computational methods to identify the entire set of OR genes in a given species, and have conducted comparative analyses of the OR gene repertoires of various organisms (6, 7, 10–15). These studies have showed that the numbers of OR genes are highly variable among different species. For example, higher primates including humans generally have smaller repertoires of (functional) OR genes (<400) than most other mammals (10–15). This observation would reflect the fact that higher primates rely on vision rather than olfaction, and thus their olfactory ability has retrogressed. Another feature of mammalian OR gene families is a high fraction (20–60 %) of pseudogenes in each species (13, 15). Moreover, evolutionary analyses revealed that the numbers of OR genes have dynamically changed during evolution due to frequent gene duplications and pseudogenization events (6, 12, 13, 15).

Here are described details of the methods to identify OR genes from mammalian genomes. These methods are also applicable and extensible to reptiles and birds, because these species have the same groups of OR genes as mammals. However, to identify OR genes from the genomes of amphibians or fish, minor methodological modifications are necessary because of the higher sequence diversity in their OR genes compared with those of mammals.

2 Materials

1. The genome sequences of various mammalian species can be downloaded from the Web sites of the University of California Santa Cruz (UCSC; genome.ucsc.edu), Ensembl Genome Browser (ensembl.org), and the Genome Sequencing Center at Washington University School of Medicine (genome.wustl.edu), the Broad Institute (www.broadinstitute.org), and others.
2. The latest version of mammalian OR gene repertoires, which are used as queries for the TBLASTN searches, can be obtained from Niimura and Nei (13), Go and Niimura (14), and Matsui et al. (15). Other databases such as GenBank (www.ncbi.nlm.nih.gov/genbank) and Olfactory Receptor Database (ORDB; senselab.med.yale.edu/ordb) (see Chapter 1) (16) also contain many OR gene sequences.

3 Methods

The strategy to identify OR genes from a given genome sequence is illustrated in Fig. 1. Here are classified OR genes identified from the genome sequences into three categories, functional genes, truncated genes, and pseudogenes. A functional gene is defined as an intact sequence that potentially encodes a functional OR, while a pseudogene is defined as a sequence containing interrupting stop codons, frameshifts, and/or gaps within conserved regions. A truncated gene is a partial intact sequence containing a part of OR gene sequence and is located at the contig end.

A truncated gene is either a functional gene or a pseudogene. If the genome sequencing is completed, it can be classified as either. For a low-coverage genome, the lengths of contigs tend to be short due to incomplete assembly. Thus, the number of truncated genes becomes large. The distinction between truncated genes and pseudogenes is necessary for estimating the fraction of pseudogenes in a given species, because the number of functional OR genes is underestimated if only intact genes are regarded as functional (13, 15).

3.1 Identification of Functional OR Genes

1. The first process to identify OR genes is to conduct TBLASTN (17) searches against a given genome sequence using known OR genes as queries. For query sequences, use a variety of known OR genes (see Note 1). The *E*-value should be 1e–10 for the identification mammalian genes as ORs.
2. Because multiple sequences are used as queries, many query sequences may hit the same genomic region. The genomic

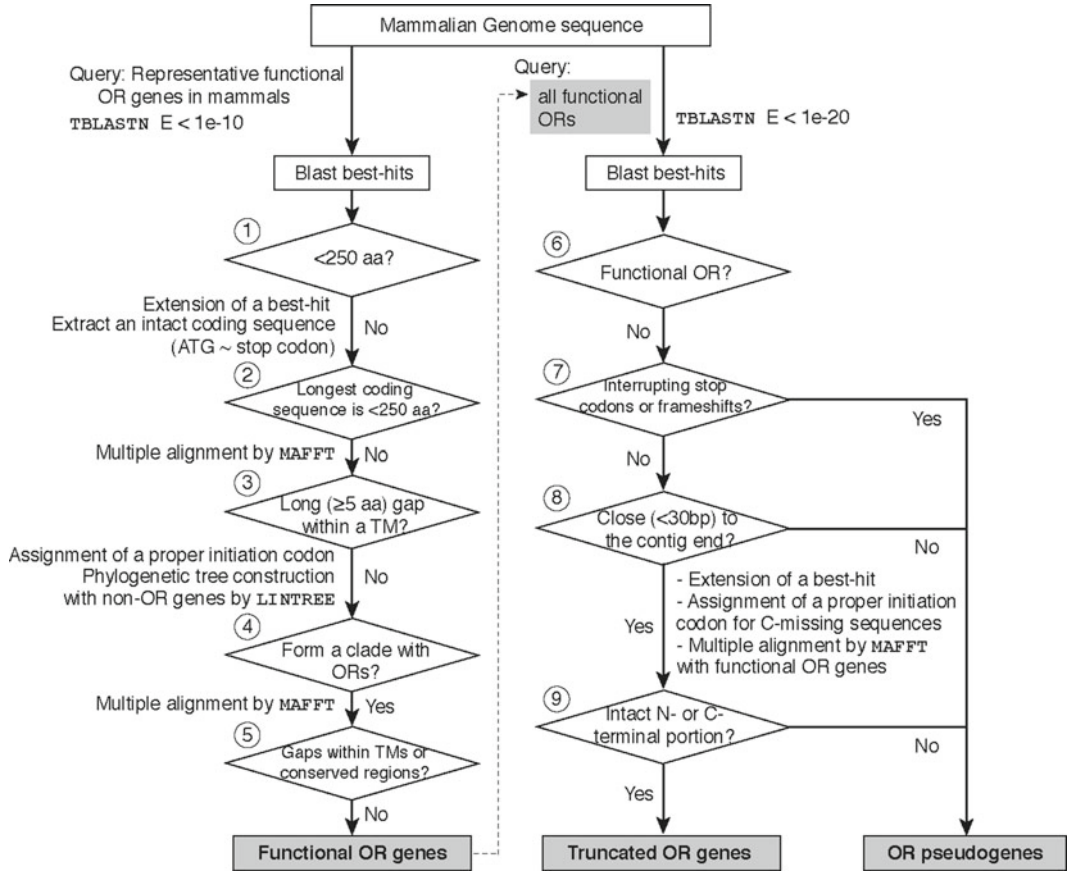

Fig. 1 Flowchart for the identification of functional OR genes, truncated OR genes, and OR pseudogenes from mammalian genome sequences

sequence that corresponds to a query showing the lowest *E*-value is the "best-hit" (Fig. 2a). Extract all the best-hit sequences from the genome.

3. Functional OR genes are identified from the best-hit sequences obtained above. To exclude nonfunctional OR genes or non-OR genes from the best-hits, the following criteria are used. First, discard the best-hit sequences that are shorter than 250 amino acids (criterion 1 in Fig. 1; see Note 2). Sometimes a query hits a genomic region that is separated into two or more fragments due to frameshifts. In such a case, examine the longest sequence among the fragments. Discard the best-hit if its length is less than 250 amino acids.
4. For each of the best-hit sequences, extend it in both directions along the genome sequence. Then, extract the longest intact coding sequence starting from an ATG codon and ending with a stop codon (TAA, TAG, or TGA) without any interrupting stop codons (Fig. 2b). If the length of the intact coding

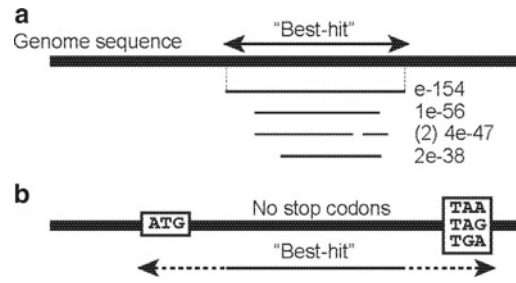

Fig. 2 (**a**) Best-hit sequence for a certain genomic region and (**b**) extension of a best-hit sequence along the genome

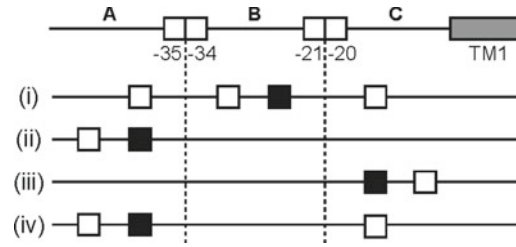

Fig. 3 Assignment of the initiation codon. The N-terminal tail of an OR gene is divided into three regions, (A) (the position −35 and its upstream), (B) (between positions −34 and −21), and (C) (the position −20 and its downstream). *Squares on lines* (*i*)–(*iv*) represent methionines encoded by ATG codons. A *black square* should be chosen as the initiation codon for each of the cases (*i*)–(*iv*)

sequence is less than 250 amino acids, discard it (criterion 2 in Fig. 1). If an extended sequence contains unknown amino acids due to the presence of undetermined nucleotides, the sequence should also be discarded.

5. Construct a multiple alignment from the remaining sequences by using the program E-INS-i in MAFFT (18) (see Note 3). Then, assign the positions of seven transmembrane (TM) regions according to Man et al. (19). If a given sequence contains a gap of five or more amino acids within at least one of the TM regions, exclude it (criterion 3 in Fig. 1).
6. In this process, for a sequence containing two or more ATG codons in the N-terminal tail (the upstream of the first TM region), the ATG codon located at the most appropriate position is chosen as the initiation codon. First, carry out a multiple alignment once again using the remaining sequences after process 5. The initiation codon is chosen according to the following criteria (Fig. 3). The N-terminal tail of a sequence is divided into three regions: the region of the position −35 and its upstream (region A: Here the amino acid position is indicated as the relative position to the boundary between the N-terminal

tail and the first TM region), the region between the positions −34 and −21 (region B), and the region of the position −20 and its downstream (region C). (1) When at least one ATG codon is present within the region B, choose the most downstream one in the region B. (2) When ATG codons are present only within the region A, choose the most downstream one. (3) When ATG codons are present only within the region C, choose the most upstream one. (4) When ATG codons are present both in region A and region C but not in region B, choose the most downstream one in the region A. The reason for using these criteria is that the length of the N-terminal tail is between 21 and 34 amino acids for most known functional OR genes.

7. Construct a phylogenetic tree for the remaining sequences together with several non-OR GPCR genes as the outgroup using the neighbor-joining (NJ) method (20) by the program LINTREE (21) (see Note 4). Use Poisson correction distances after every alignment gap is eliminated. The following genes are used as the outgroup sequences: alpha-1A-adrenergic receptor isoform 1 (GenBank protein id, NP_000671), beta-1-adrenergic receptor (NP_000675), adenosine A2b receptor (NP_000667), histamine receptor H2 (NP_071640), 5-hydroxytryptamine (serotonin) receptor 1B (NP_000854), 5-hydroxytryptamine (serotonin) receptor 1F (NP_000857), 5-hydroxytryptamine (serotonin) receptor 6 (NP_000862), galanin receptor 1 (NP_001471), somatostatin receptor 4 (NP_001043), GPCR148 (AY569570-1), and putative GPCR in mice (AJ401359-1). The reason for using these genes as the outgroup is that they are relatively close to OR genes among the genes belonging to the rhodopsin-like GPCR superfamily (4).
8. Remove non-OR genes on the basis of the phylogenetic tree constructed above (criterion 4 in Fig. 1). In the phylogenetic tree, OR genes form a monophyletic clade supported with a high bootstrap value. Therefore, non-OR genes are easily distinguishable from OR genes. When a given sequence is located out of the OR gene clade in a phylogenetic tree, it should be discarded.
9. Construct a phylogenetic tree once again using the remaining sequences after the process 8 together with non-OR genes as the outgroup (see Note 4). As for the outgroup, use the same genes as those in the process 7.
10. Classify OR genes into Class I and Class II on the basis of the phylogenetic trees constructed above (see Note 5).
11. Construct multiple alignments for Class I and Class II genes separately by using the program E-INS-i in MAFFT (18) (see Note 3).
12. Delete the sequences that have gaps at conservative amino acid sites. In Fig. 4, an amino acid site with an alphabet indicates

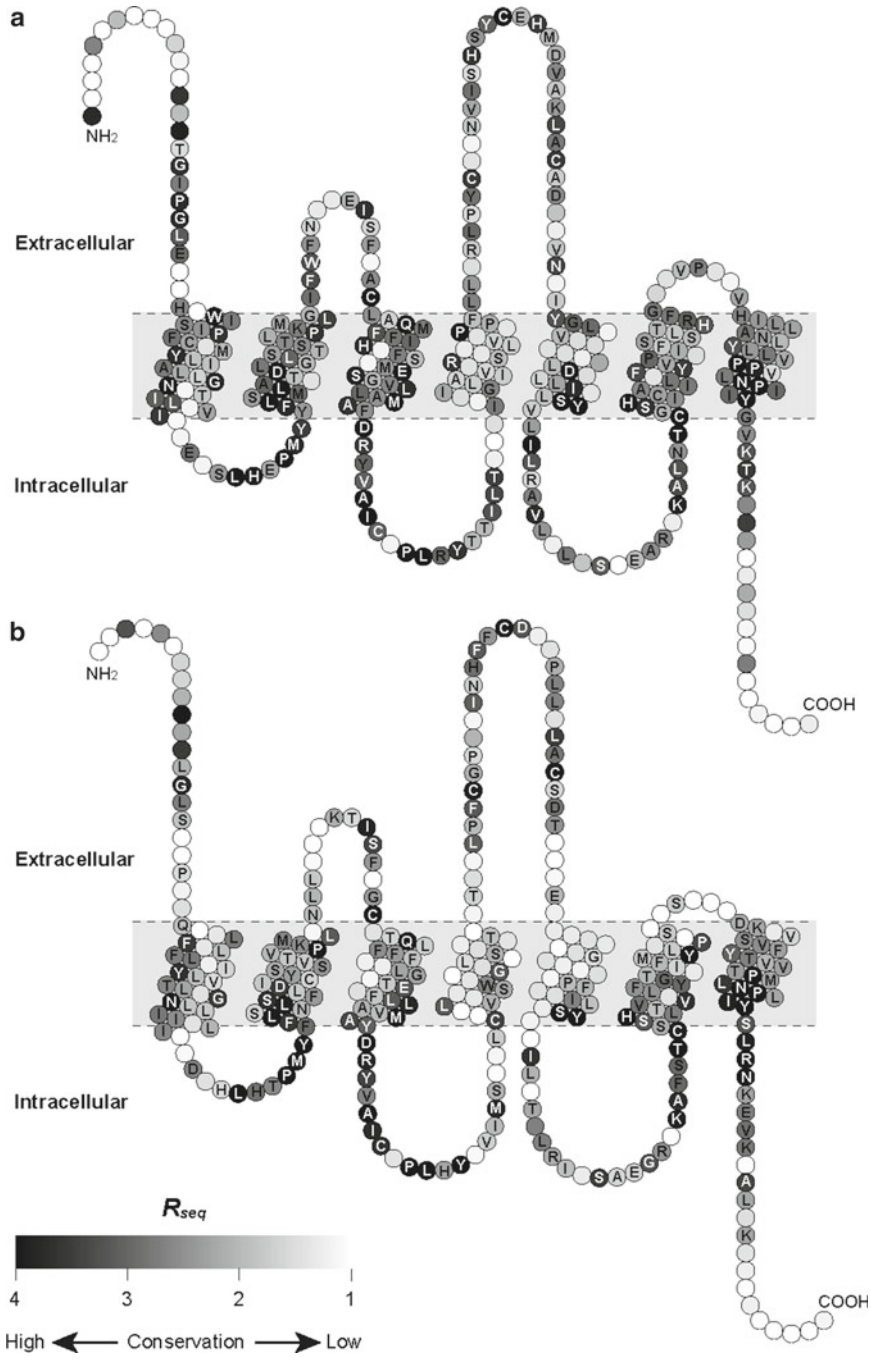

Fig. 4 Conservative amino acid sites for Class I (**a**) and Class II (**b**) OR genes. The *darkness* of a *circle* represents an extent of amino acid conservation (R_{seq}) at each position. At the amino acid site with the R_{seq} value of 1.5 or larger, the most frequent amino acid at the position is shown in the *single letter code*

the position that meets both of the following conditions: (1) Gaps are present at the position in <1 % of the entire sequences examined (Here, all functional OR genes from five primate species: human, chimpanzee, orangutan, macaque, and marmoset (15) are examined.) (2) The sequence conservation at the position in an alignment, R_{seq}, is larger than 1.5. R_{seq} is defined by the following formula (22):

$$R_{seq} = \log_2 20 - \left(-\sum_{n=1}^{20} p_n \log_2 p_n \right)$$

Here, p_n is the observed frequency of amino acid n at a given position. If a sequence contains three or more gaps at the positions indicated with alphabets in Fig. 4 and/or those within TM regions, such sequences should be excluded (criterion 5 in Fig. 1). Class I and Class II genes are examined separately. After this process, the remaining sequences are regarded as functional OR genes.

3.2 Identification of Truncated Genes and Pseudogenes

1. Conduct TBLASTN (17) searches against genome sequences using all functional OR genes in the species identified above as queries with the *E*-value below 1e–20 (see Note 6).
2. Extract all the best-hit sequences, as described in step 2 in the previous section.
3. Exclude all functional OR genes identified in the previous section (criterion 6 in Fig. 1). All the remaining sequences are regarded as pseudogenes or truncated genes.
4. Extract the best-hit sequences that meet both the following conditions: (1) There are no interrupting stop codons and frameshifts (criterion 7 in Fig. 1). (2) The distance between the end of the sequence and the end of the contig containing the sequence is less than 30 base pairs (criterion 8 in Fig. 1).
5. Classify the remaining sequences after the process 4 into three categories, C-missing, N-missing, and NC-missing (Fig. 5). For a C-missing sequence, the upstream portion of the best-hit is present in the contig examined, while its downstream is missing from the contig. Conversely, for an N-missing sequence, the downstream portion of the best-hit is present in the contig, whereas its upstream portion is missing. An NC-missing sequence is one in which both upstream and downstream portions are missing. NC-missing sequences arise only on a contig shorter than the length of an OR gene.
6. For a C-missing sequence, extend it along the genome sequence to extract a sequence from an ATG codon to the

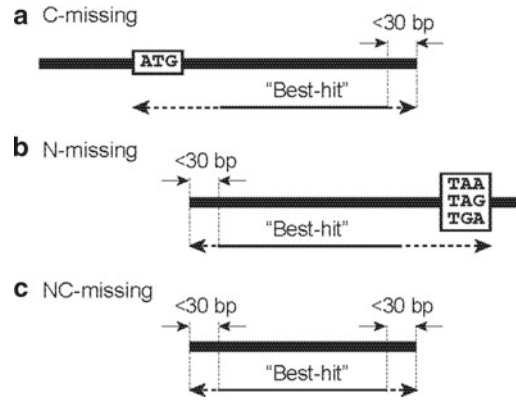

Fig. 5 Extension of a best-hit for C-missing (**a**), N-missing (**b**), and NC-missing sequences (**c**)

most downstream in-frame codon. As for the ATG codon, take the most upstream in-frame one without any interrupting stop codons (Fig. 5a).

7. For an N-missing sequence, extend it to take a sequence from the most upstream in-frame codon to the stop codon (Fig. 5b).
8. For an NC-missing sequence, extend it to take the longest sequence from the most upstream in-frame codon to the most downstream one (Fig. 5c).
9. Construct a multiple alignment using the extended N-missing, C-missing, and NC-missing sequences together with functional OR genes identified above by the program E-INS-i in MAFFT (18).
10. For each of the N-missing sequences, take the most proper ATG codon as the initiation codon. The criterion for choosing the initiation codon is the same as that in step 5 of the previous section.
11. Construct a multiple alignment once again by using all sequences resulting from step 10.
12. Exclude the sequences that do not have intact N- or C-terminal portions of a gene (criterion 9 in Fig. 1). For a C-missing sequence, discard it if there are five or more gaps in the first TM region in the alignment. For an N-missing sequence, discard it if there are five or more gaps in the seventh TM region. All the remaining sequences after this exclusion process are regarded as truncated genes.
13. Exclude all the truncated genes from the sequences obtained in step 3. All the remaining sequences are regarded as pseudogenes.

4 Notes

1. To save computational time, it is preferable to exclude highly similar sequences from a query gene set. To identify mammalian OR genes, 85 functional OR genes in human and mouse that show a 50 % or less amino acid identity to one another were used as query sequences.
2. All known functional OR genes used are more than 270 amino acids long. Therefore, the cutoff length of 250 amino acids is sufficiently shorter than any known functional OR genes. A conservative cutoff length is used to avoid incorrect exclusion of functional genes.
3. If the number of sequences is significantly higher than 400, it is preferable to separate them into several groups each of which contains less than 400 sequences, to reduce the computational time for the construction of multiple alignments.
4. If the number of sequences is higher than 200, it is preferable to separate them into several groups each of which contains less than 200 sequences, to reduce the computational time for the construction of phylogenetic trees and the calculation of bootstrap values.
5. In most cases, the Class I and Class II gene clades are supported with relatively high bootstrap values in a phylogenetic tree. However, if the distinction between Class I and Class II clades is unclear, add several mammalian Class I and Class II OR genes that are present in the database to the dataset and construct a phylogenetic tree once again.
6. The reason for using the cutoff *E*-value of 1e–20 is as follows. First, the *E*-value for a best-hit to the genomic region encoding a non-OR gene is 1e–18 or larger. Second, all best-hit sequences with the *E*-value below 1e–20 obtained by OR gene queries are more similar to OR genes than to any known non-OR genes. Therefore, OR pseudogenes and non-OR genes are distinguishable by using the *E*-value of 1e–20.

Acknowledgments

This work was supported by grant (20770192 and 23770271) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

1. Niimura Y (2009) Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Hum Genomics 4:107–118
2. Niimura Y (2012) Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. Curr Genomics 13:103–114
3. Buck L, Axel R (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–187
4. Fredriksson R, Lagerström MC, Lundin LG et al (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63:1256–1272
5. Young JM, Shykind BM, Lane RP et al (2003) Odorant receptor expressed sequence tags demonstrate olfactory expression of over 400 genes, extensive alternate splicing and unequal expression levels. Genome Biol 4:R71
6. Niimura Y, Nei M (2005) Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc Natl Acad Sci USA 102:6039–6044
7. Niimura Y (2009) On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol 1:34–44
8. Glusman G, Bahar A, Sharon D et al (2000) The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm Genome 11:1016–1023
9. Nei M, Niimura Y, Nozawa M (2008) The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet 9:951–963
10. Niimura Y, Nei M (2003) Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci USA 100:12235–12240
11. Niimura Y, Nei M (2005) Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. Gene 346:13–21
12. Niimura Y, Nei M (2005) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. Gene 346: 23–28
13. Niimura Y, Nei M (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2:e708
14. Go Y, Niimura Y (2008) Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Mol Biol Evol 25:1897–1907
15. Matsui A, Go Y, Niimura Y (2010) Degeneration of olfactory receptor gene repertories in primates: no direct link to full trichromatic vision. Mol Biol Evol 27:1192–1200
16. Crasto C, Marenco L, Miller PL et al (2002) Olfactory receptor database: a metadata-driven automated population from sources of gene and protein sequences. Nucleic Acids Res 30:354–360
17. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
18. Katoh K, Kuma K, Toh H et al (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–518
19. Man O, Gilad Y, Lancet D (2004) Prediction of the odorant binding site of olfactory receptor proteins by human-mouse comparisons. Protein Sci 13:240–254
20. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425
21. Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of molecular clock and linearized trees. Mol Biol Evol 12:823–833
22. Crooks GE, Hon G, Chandonia JM et al (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190