Chapter 7

Identification of Chemosensory Receptor Genes from Vertebrate Genomes

Yoshihito Niimura

Abstract

Chemical senses are essential for the survival of animals. In vertebrates, mainly three different types of receptors, olfactory receptors (ORs), vomeronasal receptors type 1 (V1Rs), and vomeronasal receptors type 2 (V2Rs), are responsible for the detection of chemicals in the environment. Mouse or rat genomes contain >1,000 OR genes, forming the largest multigene family in vertebrates, and have >100 V1R and V2R genes as well. Recent advancement in genome sequencing enabled us to computationally identify nearly complete repertories of OR, V1R, and V2R genes from various organisms, revealing that the numbers of these genes are highly variable among different organisms depending on each species' living environment. Here I would explain bioinformatic methods to identify the entire repertoires of OR, V1R, and V2R genes from vertebrate genome sequences.

Key words Olfactory receptor, Vomeronasal receptor, Multigene family, Bioinformatics, Vertebrate, G protein-coupled receptor

1 Introduction

Chemical senses are essential for the survival of most animals. In vertebrates, chemical molecules in the environment are mainly detected by three different types chemosensory receptors, named olfactory receptors (ORs), vomeronasal receptors type 1 (V1Rs), and vomeronasal receptors type 2 (V2Rs) (for review, Ref. 1). All of them are G protein-coupled receptors (GPCRs), membrane proteins having seven transmembrane (TM) α-helical regions (Fig. 1). Each type of receptors is encoded by a multigene family. The three gene families share almost no sequence similarity, though their molecular structures are similar to one another.

Among the three gene families, the OR gene family is by far the largest (for review, Ref. 2). OR genes are predominantly expressed in the olfactory epithelium of the nasal cavity and were first identified from rats by Linda Buck and Richard Axel in 1991 [3]. The genomes of many mammalian species harbor ~1,000 or more

Kazushige Touhara (ed.), *Pheromone Signaling: Methods and Protocols*, Methods in Molecular Biology, vol. 1068,
DOI 10.1007/978-1-62703-619-1_7,

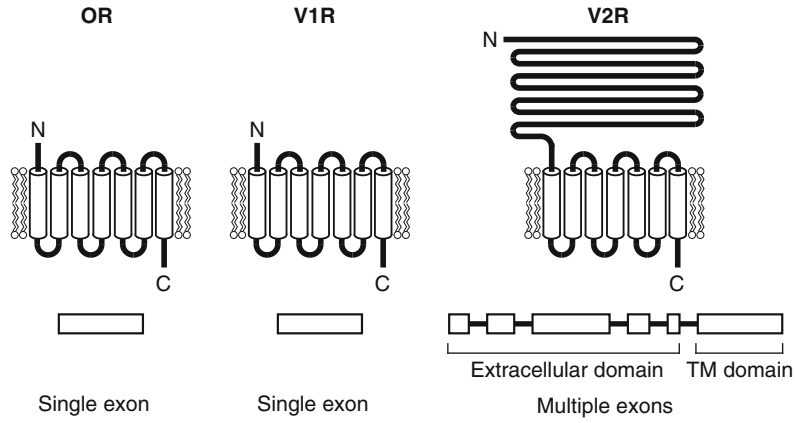

Fig. 1 Structures of ORs, V1Rs, and V2Rs. Membrane topologies and exon–intron structures of genes are shown

OR genes, which comprise 4–5 % of their proteomes. An OR is ~310 amino acids long on average, and typically the gene does not have any introns in its coding region. OR genes are present in all vertebrate species. Fishes have much smaller repertoires of OR genes (~100) than mammals. However, despite smaller OR gene repertoires in fishes, OR gene sequences in fishes are more diverse than those in mammals [4, 5]. Amphibians and fishes have some additional groups of OR genes that are not present in mammals.

V1R and V2R genes were discovered in 1995 [6] and 1997 [7–9], respectively. They are expressed in vomeronasal organs in mammals and are involved in pheromone detection. Fishes do not have a discrete vomeronasal organ, but they do have both V1R and V2R genes, which are expressed in olfactory epithelium in fishes [10, 11]. V1R genes are intronless like OR genes, whereas V2R genes consist of multiple exons (Fig. 1). A V2R gene is characterized by a long N-terminal extracellular domain, which is encoded by several exons, while the TM domain typically corresponds to a single exon.

Recently the whole genome sequences became available from diverse organisms, which enabled us to identify nearly complete repertoires of chemosensory receptor genes in a given species. Comparative genomic studies revealed that the repertoires of chemosensory receptor genes are highly variable among different species (Refs. 4, 5, 12–14 for OR genes, Refs. 10, 15–18 for V1R genes, and Refs. 17, 19, 20 for V2R genes). In this chapter, I describe the methods to identify OR, V1R, and V2R genes from vertebrate genomes. OR and V1R genes do not have any introns; therefore, detection of these genes from genome sequences is relatively easy. However, because the numbers of the genes are huge, development of computation methods is requisite for the identification of the entire gene repertoires. On the other hand, identification of V2R genes is more difficult due to their complex gene structures.

2 Materials

1. The following computer programs need to be installed: BLAST (http://blast.ncbi.nlm.nih.gov/; Ref. 21) for homology searches, MAFFT (http://mafft.cbrc.jp/alignment/software/; Ref. 22) for constructing multiple alignments, and LINTREE (http://www.personal.psu.edu/nxm2/software.htm; Ref. 23) for phylogenetic tree construction. Moreover, the HMMER package (http://hmmer.janelia.org; Ref. 24) and the Wise2 package (http://www.ebi.ac.uk/Tools/Wise2/; Ref. 25) are used for the identification of V2R genes.
2. The genome sequences of various vertebrate species can be downloaded from the Web sites of the University of California Santa Cruz (http://genome.ucsc.edu), Ensembl Genome Browser (http://ensembl.org), the Genome Sequencing Center at Washington University School of Medicine (http://genome.wustl.edu), the Broad Institute (http://www.broadinstitute.org), etc.

3 Methods

Here I would explain the method for identifying OR genes in some detail. Figure 2 illustrates the flowchart of OR gene identification. OR genes are classified into three categories: intact genes, truncated genes, and pseudogenes. An intact gene putatively encodes a functional OR. 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 encoding a part of OR and is located at the contig end. It is either a functional gene or a pseudogene. Therefore, the fraction of pseudogenes in a given species is overestimated if only intact genes are considered to be functional. This effect is critical for the species with low-coverage genome data, which contain many short contigs [5, 12, 14]. For this reason, some caution is necessary to estimate the fraction of pseudogenes.

The methods to identify V1R and V2R genes were described in Refs. 15, 16 and Refs. 19, 20, 26, respectively. The methods explained here are based on these articles with some modification.

3.1 Identification of OR Genes

1. Conduct TBLASTN searches [21] with a cutoff *E*-value of 1e−5 against a given genome sequence using known OR genes as queries (*see* **Note 1**). Use the option "-F F" (filtering low-complexity regions is not used) (*see* **Note 2**).
2. Because multiple sequences are used as queries, a number of different queries may hit the same genomic region. In such a case, choose a BLAST hit with the lowest *E*-value (called a

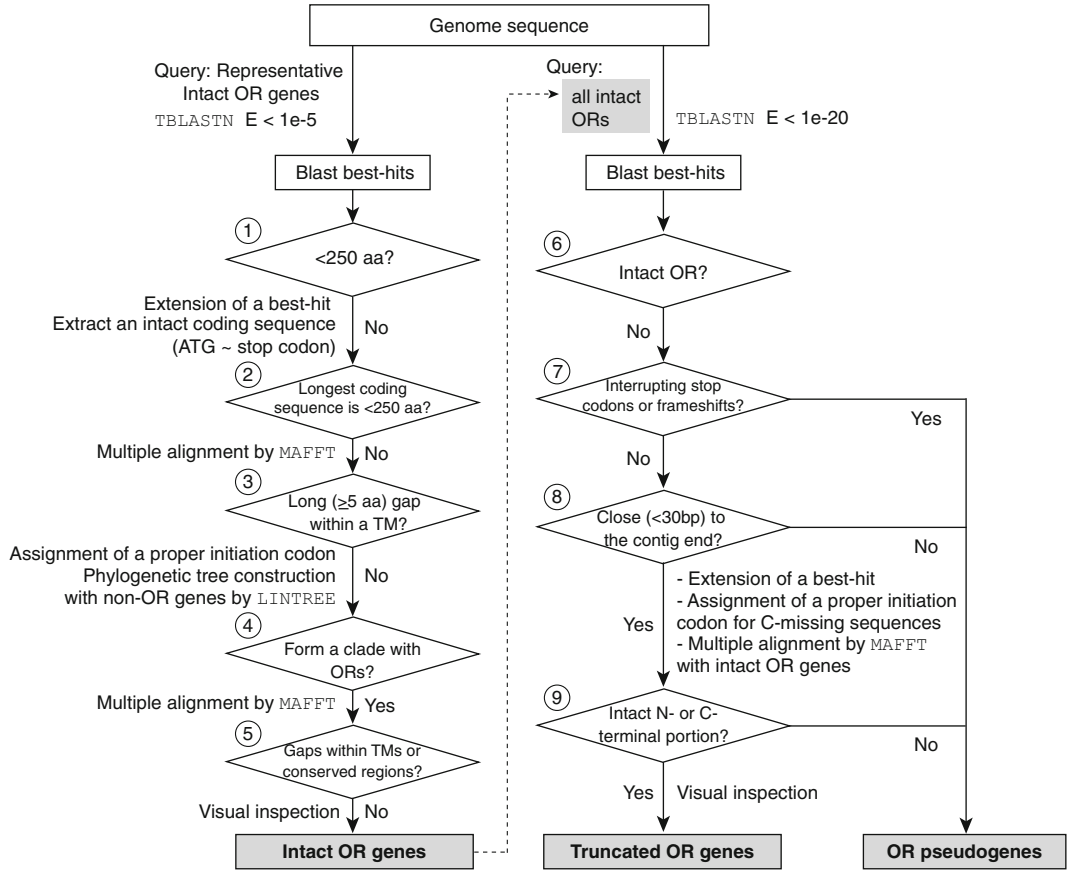

Fig. 2 Flowchart for the identification of intact OR genes, truncated OR genes, and OR pseudogenes from vertebrate genome sequences

"best-hit") among all BLAST hits to a given genomic region. Extract all best-hits from the genome sequence.

3. To identify intact OR genes from the best-hit sequences, several criteria are used. First, discard the best-hit sequences shorter than 250 amino acids (criterion 1 in Fig. 2; *see* **Note 3**).
4. For each of the best-hit sequences remaining after **step 3**, extend it to both directions along the genome sequence. Then, extract the longest coding sequence from an ATG codon to a stop codon (Fig. 3a). If the length of the sequence after extension is less than 250 amino acids, discard it (criterion 2 in Fig. 2). If the extended sequence contains undetermined nucleotides, the sequence should also be discarded.
5. Construct a multiple alignment from the remaining sequences after **step 4** by using the program MAFFT [22] (*see* **Note 4**), and assign the positions of seven TM regions according to Ref. 27.

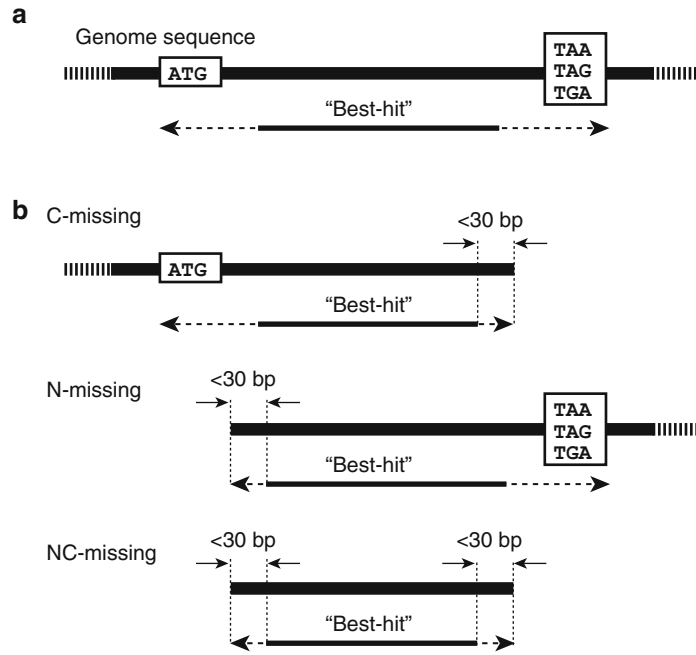

Fig. 3 (**a**) Extension of a best-hit along the genome sequence to take the longest coding sequence. (**b**) Extension of a best-hit that is located near the end of a contig. C-missing, N-missing, and NC-missing sequences are shown separately

6. If a sequence has a gap of five or more amino acids within TM regions, exclude it (criterion 3 in Fig. 2).
7. Construct a multiple alignment once again from the remaining sequences after **step 6** by using MAFFT [22] (*see* **Note 4**). Then choose the most proper ATG codon as the initiation codon in case that a sequence examined contains two or more ATG codons in the N-terminal tail region (the upstream of the first TM region; *see* **Note 5**).
8. Construct a neighbor-joining phylogenetic tree [28] using the remaining sequences after **step 7** together with several non-OR GPCR genes as the outgroup (*see* **Note 6**). Run the program njboot in LINTREE [23] with the option "-d28" (Poisson correction distance) and "-b500" (bootstrap resamplings for 500 times).
9. Eliminate non-OR genes on the basis of the phylogenetic tree constructed in **step 8** (criterion 4 in Fig. 2; *see* **Note 7**). When a given sequence is located out of the OR gene clade in the phylogenetic tree, it should be discarded (*see* **Note 8**).
10. Construct a multiple alignment from the remaining sequences after **step 9** by MAFFT [22] (*see* **Note 4**), and eliminate the sequences having gaps within TM regions or at other conserva-

tive amino acid sites by visual inspection (criterion 5 in Fig. 2). The remaining sequences are regarded as intact OR genes.

11. To identify non-intact OR genes, perform TBLASTN searches [21] against the genome sequence using all intact OR genes identified in **step 10** as queries with the *E*-value below 1e–20 (*see* **Note 9**).
12. Extract all best-hit sequences in the same way as **step 2**.
13. Exclude all of the intact OR genes obtained in **step 10** (criterion 6 in Fig. 2). All remaining sequences are regarded to be truncated genes or pseudogenes.
14. Extract the best-hit sequences that meet both of the following conditions. (1) There are no interrupting stop codons and frameshifts (criterion 7 in Fig. 2). (2) The distance between the end of the sequence and the end of the contig containing the sequence is less than 30 bp (criterion 8 in Fig. 2).
15. Classify the remaining sequences after **step 14** into three category, C-missing, N-missing, and NC-missing (Fig. 3b). For a C-missing sequence, the upstream of the best-hit is present in the contig examined, whereas its downstream is missing. Conversely, for an N-missing sequence, its downstream is present in the contig, while its upstream is missing. An NC-missing sequence lacks both upstream and downstream portions (*see* **Note 10**).
16. For a C-missing sequence, extend it along the genome sequence and extract a sequence from the most upstream ATG codon to the most downstream in-frame codon without any interrupting stop codons (Fig. 3b, top).
17. Construct a multiple alignment using the extended C-missing sequences obtained in **step 16** together with some representative intact OR genes by MAFFT [22], and choose the most proper ATG codon as the initiation codon for each sequence in the same manner as **step 7** and **Note 5**.
18. For an N-missing sequence, extend it and extract a sequence from the most upstream in-frame codon to the stop codon (Fig. 3b, middle).
19. For an NC-missing sequence, extend it and extract the longest sequence from the most upstream in-frame codon to the most downstream one (Fig. 3b, bottom).
20. Construct a multiple alignment using all of the extended C-missing, N-missing, and NC-missing sequences obtained in **steps 17–19** together with some representative intact OR genes by MAFFT [22].
21. Exclude the sequences that contain gaps within TM regions or at other conservative amino acid sites by visual inspection

(criterion 9 in Fig. 2). The remaining sequences are regarded to be truncated genes.

22. Exclude all truncated genes from the best-hit sequences obtained in **step 13**. All remaining sequences are regarded as OR pseudogenes.

3.2 Identification of V1R Genes

1. Conduct TBLASTN searches [21] with a cutoff *E*-value of 1e−5 against a given genome sequence using known V1R genes as queries. The following options should be used: "-F F" for that filtering low-complexity regions is not used and "-v 1000 -b 1000" for the number of hits reported.
2. From the results obtained in **step 1**, extract all best-hit sequences in the same manner as Subheading 3.1, **step 2**.
3. For each of the best-hit sequences, conduct BLASTP searches [21] against the nr database of GenBank to ensure that the best hit is a V1R gene. Discard the sequences showing higher similarity to non-V1R genes (*see* **Note 11**).
4. Among the remaining sequences after **step 3**, extract sequences that do not contain any interrupting stop codons or frame-shifts. For each of these sequences, extend it to both directions along the genome sequence and extract the longest coding sequence from an ATG codon to a stop codon.
5. Construct a multiple alignment from the sequences obtained in **step 4** by MAFFT [22] and choose the most proper initiation codon from each sequence.
6. Assign the location of TM regions in the multiple alignment. If a sequence contains gaps within TM regions or other highly conserved regions, it should be regarded as a pseudogene. The remaining sequences are regarded to be intact V1R genes.
7. Exclude all intact V1R genes (**step 6**) from the sequences obtained after **step 3**. The remaining sequences are regarded as V1R pseudogenes.

3.3 Identification of V2R Genes

1. The first step is to perform TBLASTN searches [21] against a given genome sequence using known V2R genes as queries. To this end, construct a multiple alignment by MAFFT [22] from the known V2R query sequences (*see* **Note 12**).
2. Trim the multiple alignment to extract the TM domain (from the first TM region to the C-terminal end) according to Ref. 8. The TM domain in each V2R gene is used as a query sequence of TBLASTN searches (*see* **Note 13**).
3. Conduct TBLASTN searches [21] with a cutoff *E*-value of 1e−5 using the TM domain of known V2R genes as queries. The following options should be used: "-F F" for that filtering low-complexity regions is not used and "-v 1000 -b 1000" for the number of hits reported.

4. From the results obtained in **step 3**, extract all best-hit sequences in the same manner as Subheading 3.1, **step 2**.
5. For each of the best-hit sequences, extract the genomic sequence together with 200 kb upstream and 1 kb downstream regions (*see* **Note 14**).
6. Construct a profile Hidden Markov Model (HMM) from the alignment generated in **step 1**. (Here use the alignment including the entire coding region rather than a TM domain). Run the program hmmbuild in the HMMER package, version 2.3.2 [24] for the MAFFT output file with options "--fast --gapmax 0.95 -s" [20]. Then run the hmmcalibrate program subsequently.
7. Align each of the elongated best-hit sequences obtained in **step 5** with a profile HMM created in **step 6** by the program genewisedb in the Wise2 package [25]. Use the following options: "-splice flat -cut 20 -alg 623 -aalg 623 -pretty -para -pseudo -genes -sum -cdna -trans -gff -gener" [20].
8. When two neighboring elongated best-hit sequences (from **step 5**) are overlapped, the same genomic region may be detected as a result of the genewisedb searches. In such cases, extract only one genewisedb hit showing the highest score among all overlapping hits.
9. For each of the genewisedb hits obtained in **step 8**, conduct BLASTP searches [21] against the nr database of GenBank to ensure that the best hit is a V2R gene. Discard the sequences showing higher similarity to non-V2R genes (*see* **Note 15**). All remaining sequences are regarded to be V2R genes.
10. Discard the sequences (1) that contain interrupting stop codons or frameshifts, (2) that have gaps within TM regions or other highly conserved regions, and (3) that are shorter than 750 amino acids. The remaining sequences are regarded to be intact V2R genes.
11. Exclude all intact V2R genes (**step 10**) from the sequences obtained after **step 9**. The remaining sequences are regarded as V2R pseudogenes.

4 Notes

1. As for query sequences, an OR gene set with sufficient sequence diversity should be used to retrieve all putative OR genes from the genome. However, currently thousands of OR genes are available in the databases; therefore, to reduce a computational time, highly similar sequences should be eliminated from the queries. To this aim, classify candidate query genes into groups using a given sequence similarity cutoff, e.g., 50 % amino acid

identity, and choose one representative sequence from each group. To examine amniote (mammalian, avian, and reptilian) genomes, human and/or mouse OR genes can be used as queries. On the other hand, to investigate amphibian or fish genomes, OR genes from fishes (e.g., zebrafish) and/or frogs should be used as queries, because amphibian and fish OR genes are more diverse than amniote OR genes [5]. OR gene sequences in mammals and other vertebrates can be obtained from Refs. 12, 14, 29 and Ref. 5, respectively.

2. If the number of BLAST hits is expected to large, e.g., when mammalian genome sequences are examined, "-v" and "-b" options should also be used to change the number of hits and alignments shown.
3. The cutoff length of 250 amino acids is sufficiently shorter than that of any known functional OR genes.
4. When the number of sequences is large (e.g., >400), it is better to separate them into several groups to reduce a computational time.
5. For most of the known functional OR genes in mammals, the length of the N-terminal tail is between 21 and 34 amino acids. Therefore, to choose the initiation codon, the following criteria can be used. If an ATG codon is present in the region (named "region A") between the positions -34 and -21 (here the amino acid position is indicated as the relative position to the boundary between the N-terminal tail and the first TM region), choose the one as the initiation codon. In case that two or more ATG codons are present within the region A, choose the most downstream one among them. If ATG codons are not present in the region A, choose the closest one to the region A.
6. The following genes can be 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), and somatostatin receptor 4 (NP_001043). These genes are relatively close to OR genes among the genes belonging to the rhodopsin-like GPCR superfamily [30].
7. When the number of sequences is large (e.g., >200), it is better to separate them into several groups to reduce a computational time.
8. In a phylogenetic tree, OR genes form a monophyletic clade with a high bootstrap support [5]. Therefore, non-OR genes are easily distinguishable from OR genes.

9. The reason for using the cutoff *E*-value of 1e–20 is as follows. First, the *E*-value of a best-hit to the genomic region corresponding to 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, non-intact OR genes can be distinguished from non-OR genes by using the cutoff *E*-value of 1e–20.
10. Note that NC-missing sequences are found only on a contig shorter than the length of an OR gene (~930 bp).
11. This step is necessary to exclude some other receptors (e.g., T2R taste receptors) that are homologous to V1Rs.
12. A V2R gene set with sufficient sequence diversity should be used as query sequences. These sequences are also used to construct a profile HMM (*see* **step 6**). To search for mammalian V2R genes, for example, 75 intact V2R genes in mice with the prefix "mouseMay04V2R" in Ref. 20 can be used as queries. Fish V2R gene sequences are available in Ref. 19.
13. The N-terminal extracellular domain shows a higher extent of sequence diversity [26]; therefore, if N-terminal extracellular domain is used as a query, homology searches give a large number of non-V2R hits. For this reason, it is better to use only a TM domain rather than the entire sequence of a V2R gene as a query.
14. For extracting the entire coding exons of a V2R gene, it is necessary to examine a long upstream genomic sequence, because the N-terminal extracellular domain is encoded by multiple exons (Fig. 1). (Note that a query sequence for homology searches contains only a TM domain.) The reason for using the 200 kb limit is that it is longer than the genomic extent of all previously described V2R genes [20].
15. This step is necessary, because some other receptors (e.g., calcium-sensing receptors and T1R taste receptors) are known to be homologous to V2Rs.

Acknowledgments

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

References

1. 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
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. Niimura Y, Nei M (2005) Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc Natl Acad Sci USA 102:6039–6044
5. 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
6. Dulac C, Axel R (1995) A novel family of genes encoding putative pheromone receptors in mammals. Cell 83:195–206
7. Herrada G, Dulac C (1997) A novel family of putative pheromone receptors in mammals with a topographically organized and sexually dimorphic distribution. Cell 90:763–773
8. Matsunami H, Buck LB (1997) A multigene family encoding a diverse array of putative pheromone receptors in mammals. Cell 90:775–784
9. Ryba NJ, Tirindelli R (1997) A new multigene family of putative pheromone receptors. Neuron 19:371–379
10. Saraiva LR, Korsching SI (2007) A novel olfactory receptor gene family in teleost fish. Genome Res 17:1448–1457
11. Hashiguchi Y, Nishida M (2005) Evolution of vomeronasal-type odorant receptor genes in the zebrafish genome. Gene 362:19–28
12. Niimura Y, Nei M (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2:e708
13. Hayden S, Bekaert M, Crider TA et al (2010) Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res 20:1–9
14. 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
15. Grus WE, Shi P, Zhang YP et al (2005) Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. Proc Natl Acad Sci USA 102:5767–5772
16. Young JM, Kambere M, Trask BJ et al (2005) Divergent V1R repertoires in five species: Amplification in rodents, decimation in primates, and a surprisingly small repertoire in dogs. Genome Res 15:231–240
17. Shi P, Zhang J (2007) Comparative genomic analysis identifies an evolutionary shift of vomeronasal receptor gene repertoires in the vertebrate transition from water to land. Genome Res 17:166–174
18. Young JM, Massa HF, Hsu L et al (2010) Extreme variability among mammalian V1R gene families. Genome Res 20:10–18
19. Hashiguchi Y, Nishida M (2006) Evolution and origin of vomeronasal-type odorant receptor gene repertoire in fishes. BMC Evol Biol 6:76
20. Young JM, Trask BJ (2007) V2R gene families degenerated in primates, dog and cow, but expanded in opossum. Trends Genet 23:212–215
21. 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
22. 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
23. Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of molecular clock and linearized trees. Mol Biol Evol 12:823–833
24. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195
25. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995
26. Yang H, Shi P, Zhang YP et al (2005) Composition and evolution of the V2r vomeronasal receptor gene repertoire in mice and rats. Genomics 86:306–315
27. 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
28. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425
29. Go Y, Niimura Y (2008) Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Mol Biol Evol 25:1897–1907
30. 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