SCIENCE CHINA
Life Sciences

•RESEARCH PAPER•

December 2022 Vol.65 No.12: 2539–2551
https://doi.org/10.1007/s11427-021-2081-6

Building the Chordata Olfactory Receptor Database using more than 400,000 receptors annotated by Genome2OR

Wei Han[1,2,3,4], Yiran Wu[1], Liting Zeng[1,2,3,4] & Suwen Zhao[1,2*]

[1] *iHuman Institute, ShanghaiTech University, Shanghai 201210, China;*
[2] *School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China;*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China;*
[4] *Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China*

Received January 8, 2022; accepted January 20, 2022; published online June 10, 2022

Olfactory receptors are poorly annotated for most genome-sequenced chordates. To address this deficiency, we developed a nhmmer-based olfactory receptor annotation tool Genome2OR (https://github.com/ToHanwei/Genome2OR.git), and used it to process 1,695 sequenced chordate genomes in the NCBI Assembly database as of January, 2021. In total, 765,248 olfactory receptor genes were annotated, with 404,426 functional genes and 360,822 pseudogenes, which represents a four-fold increase in the number of annotated olfactory receptors. Based on the annotation data, we built a database called Chordata Olfactory Receptor Database (CORD, https://cord.ihuman.shanghaitech.edu.cn) for archiving, analysing and disseminating the data. Beyond the primary data, we offer derivative information, including pictures of species, cross references to public databases, structural models, sequence similarity networks and sequence profiles in the CORD. Furthermore, we did brief analyses on these receptors, including building a huge protein sequence similarity network covering all receptors in the database, and clustering them into 20 communities, classifying the 20 communities into three categories based on their presences/absences in ray-finned fish and/or lobe-finned fish. We infer that olfactory receptors should have unique activation and desensitization mechanisms by analysing their sequences and structural models. We believe the CORD can benefit the researchers and the general public who are interested in olfaction.

G protein-coupled receptors, olfactory receptors, Genome2OR, CORD

Citation: Han, W., Wu, Y., Zeng, L., and Zhao, S. (2022). Building the Chordata Olfactory Receptor Database using more than 400,000 receptors annotated by Genome2OR. Sci China Life Sci 65, 2539–2551. https://doi.org/10.1007/s11427-021-2081-6

INTRODUCTION

Olfaction is one of the most important senses in chordates, and it plays crucial roles in finding food, avoiding danger, mating, identifying individuals, recognizing marker territories etc. (Nei et al., 2008; Niimura, 2012; Touhara and Vosshall, 2009). Most olfaction-relevant receptors in chordates are G protein-coupled receptors (GPCRs), including olfactory receptors (ORs), vomeronasal type-1 and type-2 receptors (V1Rs, V2Rs), and trace amine-associated receptors (TAARs). Among these four types of receptors, ORs form the largest family in the GPCR superfamily. The number of olfactory receptor genes (*Olfrs*) varies significantly among different species, and a large portion of *Olfrs* are pseudogenes in most species (Go and Niimura, 2008; Matsui et al., 2010; Niimura, 2009b; Niimura and Nei, 2003; Niimura and Nei, 2005a; Niimura and Nei, 2005b; Niimura and Nei, 2007). Extensive phylogenetic analyses have suggested that the number of *Olfr* gains and losses is enormous (Liu et al., 2019; Niimura et al., 2014; Niimura and Nei, 2007), and that *Olfrs* are dynamically changing, depending on the ecological niche of each species occupies

*Corresponding author (email: zhaosw@shanghaitech.edu.cn)

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

(Niimura, 2009a).

Olfrs are poorly annotated, primarily due to lacking of a powerful tool and consequently without a comprehensive database. As of January 2021, the whole genomes of 1,695 chordates have been archived in the NCBI Assembly database. However, the annotations of *Olfrs* in these genomes vary dramatically. While the *Olfrs* of human and a few model organisms are well annotated (Liu et al., 2011; Olender et al., 2013; Skoufos et al., 1999), those of most other species are poorly annotated. Most strikingly, 83% of genome-sequenced chordates do not even have a single *Olfr* recorded in the UniProt database (Figure S1 in Supporting Information).

Additionally, most *Olfrs* are also not recorded in currently available OR-centred databases, including Olfactory Receptor Database (ORDB) (Skoufos et al., 1999), Human Olfactory Data Explorer (HORDE) (Olender et al., 2013) and ODORactor (Liu et al., 2011). Among the three databases, ORDB integrated OR research results from more than 100 labs, it is a repository of 18,735 chemosensory receptors from 70 species. The latest version of ORDB is version 6.0 released in 2015. HORDE has 6,739 functional and 4,336 pseudo *Olfrs* from 9 species, including human. For human ORs, HORDE has state-of-the-art annotations on genomic variations, classification, orthologs and other features. ODORactor is a web server of 1,516 functional and 92 pseudo *Olfrs* from 2 species, focusing on OR-odorant interactions. The three databases comprehensively record OR sequences, OR ligands, and/or olfactory pseudogenes for human and very limited other species (Table S1 in Supporting Information) (Marenco et al., 2016). However, the numbers of species recorded are all less than 80, which represents less than 5% of the species that have been genome-sequenced, leaving *Olfrs* not annotated in a much larger number of whole genome sequenced species.

While the sequencing revolution has led to the sequencing and assembly of tens of thousands of new genomes, eukaryotic genome annotation is still highly dependent on nearly the same technology that has been used during the past two decades (Salzberg, 2019). These generic genome annotation methods (Gross et al., 2007; Holt and Yandell, 2011; Sharma et al., 2016; Stanke et al., 2004) are optimized to solve challenging problems such as few and widely separated genes in the genome, and genes that are interrupted by introns in eukaryotes. It seems that none of them are optimized to annotate *Olfrs*, although in many species *Olfrs* constitute about 5% of the genomes. That leads to an unexpected fact that *Olfrs* from most sequenced chordate genomes are poorly annotated or not annotated at all.

The annotation of *Olfrs* seems a less challenging problem since the vast majority of chordate *Olfrs* have only one exon (Buck and Axel, 1991; Nef et al., 1992), with the exception of cichlids that have a small number of *Olfrs* with 2–4 exons (Azzouzi et al., 2014). This single-exon property of *Olfrs* and the conservative protein sequence patterns of ORs enable a rather straightforward annotation of *Olfrs* for a given genome. Therefore, a few researchers have designed custom annotation processes to identify *Olfrs* from sequenced genomes.

In 2005, Quignon et al. identified 1,009 *Olfrs* from dog genome and 1,493 *Olfrs* from the rat genome (Quignon et al., 2005). They generated five amino acid patterns for dog and rat respectively, using 45 already annotated full-length canine ORs and 200 rat ORs in public database. Next, they used the predefined five amino acid patterns to scan all six translation frames of CanFam1.0 and Rnor3.1 to identify *Olfrs*. Obviously, using five predefined patterns is a primitive and heuristic approach that are not systematic and robust enough to be applied to all chordates. Scanning all six translation frames of the entire genome rather than using DNA sequence profiles to locate *Olfrs* also severely limits the speed of annotation.

In 2010, Hayden et al. developed a TFASTX-based (Pearson et al., 1997) *Olfrs* annotation tool named Olfactory Receptor family Assigner (ORA, https://github.com/pseudogene/ora) (Hayden et al., 2010). The ORA performs TFASTX searches, using close to 200 representative OR protein sequences as queries to identify *Olfrs* in genomes, then these genes are assigned into 17 OR families according to their profile hidden Markov models (profile HMMs) (Eddy, 1998). ORA requires that the *Olfr* to be annotated does not exceed 40 kb otherwise the program will die, which makes it inconvenient to use.

In 2013, Niimura developed a tblastn-based method to find potential *Olfrs* in a given genome (Niimura, 2013). He successfully applied the method to quite a few species (Go and Niimura, 2008; Matsui et al., 2010; Niimura, 2009b; Niimura and Nei, 2003; Niimura and Nei, 2005a; Niimura and Nei, 2005b; Niimura and Nei, 2007). However, Niimura's code is not public. In 2015, Fan et al. reimplemented Niimura's approach and named it as ORFAM, which is available on GitHub (https://github.com/jianzuoyi/orfam). For most mammals, this method can find most of the functional *Olfrs* and pseudogenes. One drawback of this approach is that the search results are highly dependent on query sequences. It is quite common that the search results are not consistent with each other when using different sequences as queries. Although tblastn is a faster method than TFASTX (Zhang, 2001), the tblastn-based ORFAM approach is still slow, which often takes several hours or more to annotate a single chordate genome.

Clearly, in the post-genome era, a powerful tool for *Olfr* annotation and a comprehensive OR database is required. In order to annotate *Olfrs* from whole genomes in a fast, sensitive and robust manner, here we report a new method called Genome2OR. By using Genome2OR, we successfully identified 765,248 *Olfrs*, including 404,426 functional *Olfrs* and 360,822 olfactory pseudogenes, from 1,695 whole genome sequenced chordates. We built a database called the Chordata Olfactory Receptor Database (CORD) to store, organize and disseminate these data.

RESULTS

Genome2OR is a fast, sensitive, and robust *Olfr* annotation method

Genome2OR is a nhmmer-based (Wheeler and Eddy, 2013), sensitive, and fast tool we developed for batch annotation of *Olfrs* from chordate genomes. It contains five main modules: nhmmer.py, FindOR.py, IdentifyFunc.py, Batch.py, and Iteration.py (Figure 1). It exploits the improved remote DNA homolog detection power of nhmmer (Wheeler and Eddy, 2013). The key input for the Genome2OR tool is an appropriate profile HMM built from known *Olfr* DNA sequences. Considering OR coding sequences are considerably different between the seven evolutionary clades: lancelets, jawless fish, jawed fish, amphibians, reptiles, birds, and mammals, we built DNA profile HMM for each clade through a strict protocol (Figure S2 in Supporting Information, please also refer to MATERIALS AND METHODS).

The performance of Genome2OR is significantly improved, compared with ORA and ORFAM, according to our test of the three methods on a benchmark dataset containing six species (Data S1 in Supporting Information) for which HORDE holds the state-of-the-art annotation of *Olfrs*. First of all, the results of Genome2OR strongly resemble that of HORDE, with 92.3%–99.4% OR sequences identified by Genome2OR from each species can be mapped to HORDE (Figure S3 in Supporting Information). Remarkably, for each species, the precisions of Genome2OR results are consistently better than that of ORA/ORFAM. Second, Genome2OR is much faster than ORA/ORFAM, it costs ~100-fold less time than ORA/ORFAM (Figure S4 in Supporting Information). In addition, Genome2OR costs at least 10-fold less memory than ORFAM (Figure S5 in Supporting Information).

Building the CORD using *Olfrs* annotated by Genome2OR from 1,695 genome sequenced species

We successfully identified 765,248 *Olfrs*, including 404,426 functional *Olfrs* and 360,822 olfactory pseudogenes, when using Genome2OR to annotate 1,695 whole genome sequenced chordates in the NCBI Assembly database (Tables S2 and S3 in Supporting Information). To benefit the community, we built a database to store, organize and disseminate the results, and named it as CORD. Both the number of species with *Olfr* annotated and the number of functional ORs in the CORD are significantly greater than that in the UniProt database which only has ~80,000 functional OR protein sequences recorded for less than 300 species (Figure 2A and B).

Using annotated *Olfrs* in the CORD, we are able to count the number of functional ORs in species of different chordate groups (Figure 2C). Generally, mammals and amphibians have a relatively large number of ORs (median 704 and 702, respectively). Birds have the smallest number of ORs (median 32), while evolutionarily related non-bird reptiles have many more ORs (median 512), revealing that most birds may rely more on vision. Fish (including jawed fish and jawless fish, median 50 and 54, respectively) do not have many ORs, showing that aquatic life may depend less on olfaction. This is further supported by fewer number of ORs in the aquatic mammals, such as dolphins and whales (9–95 functional ORs), compared with terrestrial mammals. However, certain fish and birds can have large number of ORs.

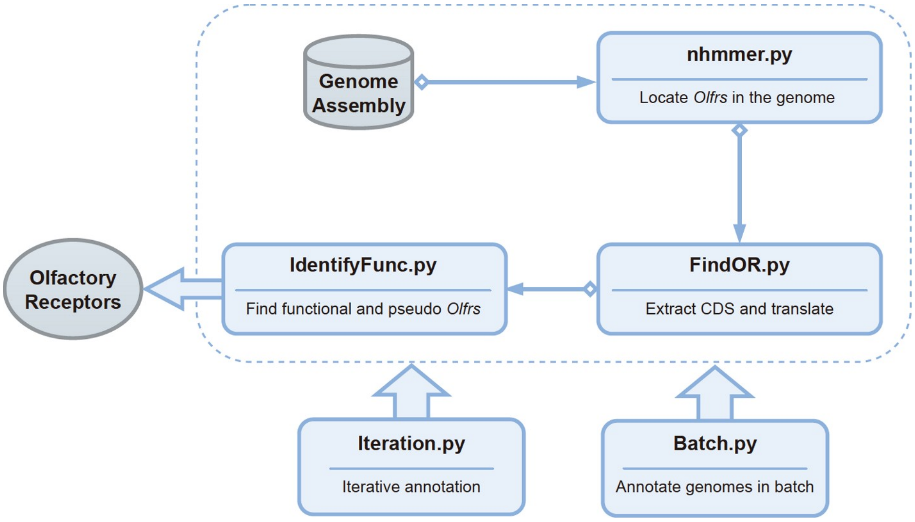

Genome Assembly
nhmmer.py
Locate Olfrs in the genome
FindOR.py
Extract CDS and translate
IdentifyFunc.py
Find functional and pseudo Olfrs
Olfactory Receptors
Iteration.py
Iterative annotation
Batch.py
Annotate genomes in batch

Figure 1 The five main python modules of Genome2OR.

For example, reedfish (*Erpetoichthys calabaricus*) have 566 functional ORs, while the northern flicker (*Colaptes auratus*) has 770 functional ORs. Why these species have so many ORs compared with their close relatives remains unknown.

In the current version of the CORD, the species with the largest number of functional ORs is the African elephant (*Loxodonta africana*), reaching 1,961, which is very close to the estimate of Niimura (2012). The species with the second largest number of functional ORs is punctate agouti (*Dasyprocta punctata*), with 1,881 functional ORs annotated.

Based on our statistics, the number of *Olfrs* (functional and total) is not correlated to the size of the genome (Figures S6 and S7 in Supporting Information). As an extreme case, the newly released Australian lungfish (*Neoceratodus forsteri*, which is the closest relatives of the land-living vertebrates) genome assembly is 37 Gb in size, the largest animal genome to date (Meyer et al., 2021). Interestingly, it only harbors 166 functional ORs according to our annotation, which is larger than 50, the median number of functional ORs in jawed fish, but still much smaller than 702, the median number of functional ORs in amphibians (Figure 2C).

Menus and contents of the CORD

The CORD has seven menus including: "Home", "Receptors", "Network", "Profiles", "BLAST", "Genome2OR", and "Help". The "Home" page provides a global search box for searching species or gene name across the CORD, and a brief statistics of the CORD. The "Receptors" menu is the data centre of the CORD. For each evolutionary clade under the "Receptors" menu, there are four hierarchical webpages, including "All species in a clade" page, "Single species" page, "All ORs in a species" page, and "Single OR" page (Figure 3). The "Network" menu provides sequence similarity network and community classification and analysis for all ORs in the CORD. All functional ORs in the CORD were clustered into 20 communities, C01 to C20. The "Profiles" menu provides the position-specific scoring matrix (PSSM, which is the frequency of occurrence for each residue at each position) and weblogo format profiles of each transmembrane helices and loops of ORs from different evolutionary clades and sequence clusters (Crooks et al., 2004). The "BLAST" menu provides sequence similarity searches for proteins and DNAs by NCBI blast+ toolkit (v2.11) (Camacho et al., 2009). The "Genome2OR" is the introduction page of the Genome2OR tool. The "Help" menu provides user access to the CORD documentation.

14 out of 20 communities only exist either in ray-finned fish or lobe-finned fish

We analysed the presence of OR communities in bony vertebrates since most ORs of ancient chordates, such as lancelets, tunicates, jawless fish, and cartilaginous fish, are too special to be clustered into communities C01–C20. According to the presence of OR communities in ray-finned fish and lobe-finned fish, the 20 communities can be divided into three groups: (i) bony vertebrate group, which includes communities C03, C09, C16, and C18–C20 that present in both ray-finned fish and lobe-finned fish; (ii) ray-finned fish group, which includes communities C08, C10, C11, C14, C15, and C17 that only present in ray-finned fish; and (iii) lobe-finned fish group, which includes communities C01, C02, C04–C07, C12, and C13 that only present in lobe-finned fish (Figure 4). The latter two groups have 14 communities, which means 14 out of 20 communities only exist either in ray-finned fish or lobe-finned fish.

Among the three community groups, the bony vertebrate group is different from the other two in that four out of six communities in this group are absent in at least one major clade (Figure 4B), while all communities in the ray-finned fish group and the lobe-finned fish group are not absent in any corresponding major clades. For example, community C03 in the bony vertebrate group exists in all clades of lobe-finned fish and ray-finned fish except teleost fish (Figure 4B). Similarly, community C18 in the vertebrate group presents in teleost fish and lungfish but not other lobe-finned fish (Figure 4B). These results showed annotation of species in minor basal clades are of great importance: if ORs in the three basal clades of ray-finned fish and lungfish were absent in database, community C03 would be classified into the lobe-finned fish group, and community C18 would be classified into ray-finned fish group.

The three community groups reflect the big picture of the community-wise birth and death of ORs during evolution, especially during the separation of lobe-finned fish and ray-finned fish. We can also use the CORD to track birth and death details of any community. For instance, community C19 (which belongs to the lobe-finned fish group) is absent in great apes (including human) but still exist in gibbons and other mammals. That is, ORs in community C19 were lost after separation of Hominidae and Hylobatidae. This seems to indicate that chordates have expanded or discarded a community in response to the change of ecological niches they live environmental changes during evolution.

Sequence and structural analyses reveal ORs have unique activation and desensitization mechanisms

We analysed the sequences and structures of 20 communities of the CORD. Since there are no experimental structures available for ORs, we used the AlphaFold2 (Jumper et al., 2021) to model representative ORs in the 20 communities (Data S2 in Supporting Information).

The weblogos of the 20 communities were used to analyse the conservation of sequences (Data S3 in Supporting

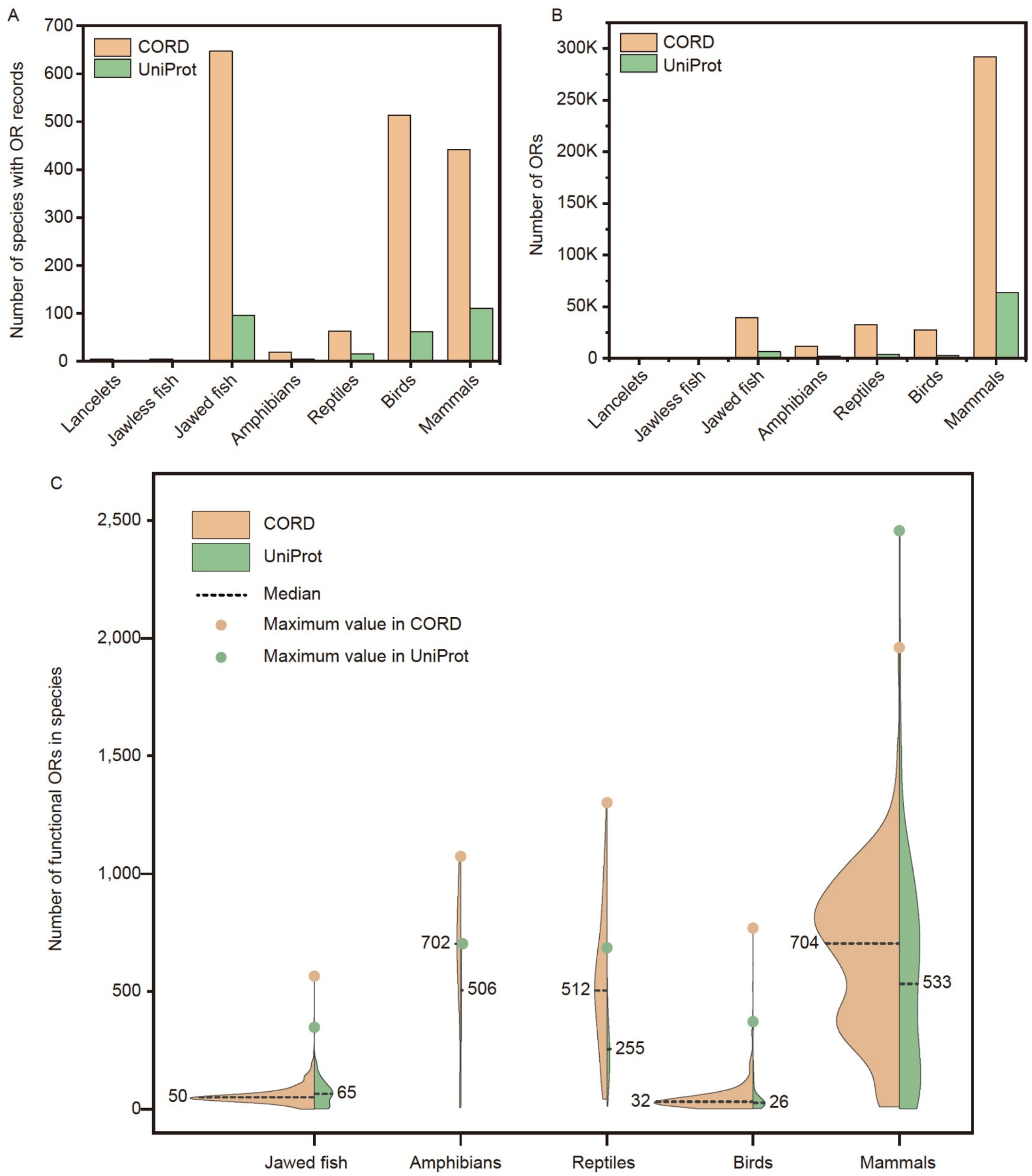
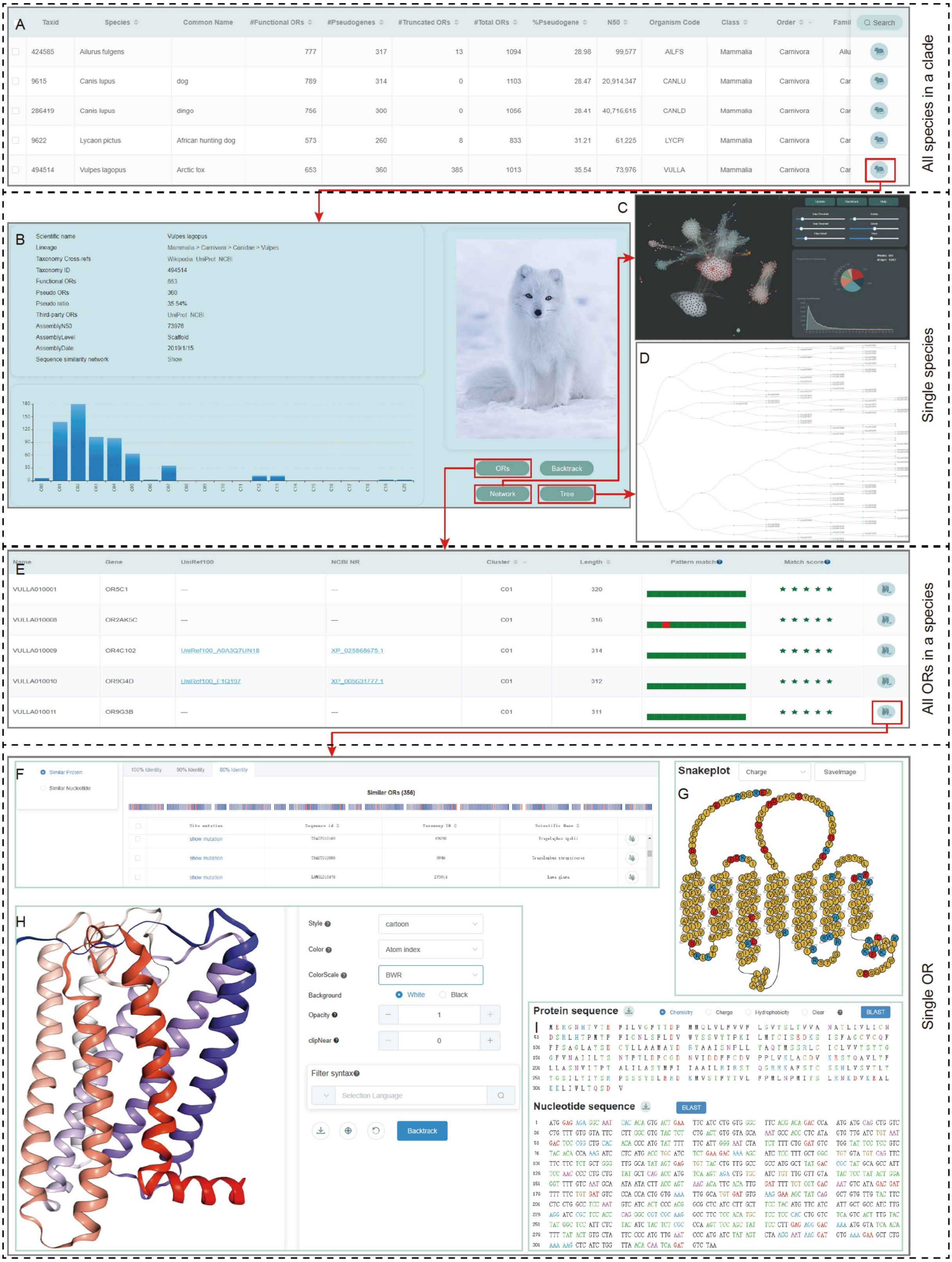


Figure 2 The number of ORs in CORD overwhelms that in UniProt. Comparison of OR sequences in CORD and UniProt (as of January 2021) in terms of (A) the number of species with functional ORs and (B) the number of functional ORs. An OR protein length filter of 250–500 amino acids was used for both databases. C, Distributions of the number of functional ORs in five evolutionary clades in CORD and UniProt. Medians and maximums are provided for each distribution. The medians changed dramatically for amphibians, reptiles, and mammals.

Information). In a community, if a residue at a position is conserved in ≥50% of the sequences in this community, this position is considered as a community conserved site. If a site is a community conserved site in at least 15 out of the 20 communities, the site is considered as a globally conserved site (Data S4 in Supporting Information). We further classify globally conserved sites into three categories, according to their degree of conservation across ≥15 communities: (i) conserved site, a site with the same residue in ≥15 communities, (ii) similar site, a site with different but physiochemically similar residues in ≥15 communities (Rives et al., 2021), and (iii) diverse site, a site with different and physiochemically dissimilar residues in ≥15 communities (Figure S8 in Supporting Information).

Based on the definitions, it is clear that the extracellular half of seven-transmembrane domain of the ORs are less conserved compared with the intracellular half. Specifically, the extracellular halves of TM4, TM5, TM6, and TM7 are very diverse (corresponding to white-coloured circles in Figure S8 in Supporting Information), reflecting the repertoire of ORs in recognizing extremely diverse odors; while the intracellular halves of TM2, TM3, TM5, TM6 and



	Taxid	Species	Common Name	#Functional ORs	#Pseudogenes	#Truncated ORs	#Total ORs	%Pseudogene	N50	Organism Code	Class	Order	Famil
□	424585	Ailurus fulgens		777	317	13	1094	28.98	99,577	AILFS	Mammalia	Carnivora	Ailu
□	9615	Canis lupus	dog	789	314	0	1103	28.47	20,914,347	CANLU	Mammalia	Carnivora	Car
□	286419	Canis lupus	dingo	756	300	0	1056	28.41	40,716,615	CANLD	Mammalia	Carnivora	Car
□	9622	Lycaon pictus	African hunting dog	573	260	8	833	31.21	61,225	LYCPI	Mammalia	Carnivora	Car
□	494514	Vulpes lagopus	Arctic fox	653	360	385	1013	35.54	73,976	VULLA	Mammalia	Carnivora	Car

Scientific name	Vulpes lagopus
Lineage	Mammalia > Carnivora > Canidae > Vulpes
Taxonomy Cross-refs	Wikipedia UniProt NCBI
Taxonomy ID	494514
Functional ORs	653
Pseudo ORs	360
Pseudo ratio	35.54%
Third-party ORs	UniProt NCBI
AssemblyN50	73976
AssemblyLevel	Scaffold
AssemblyDate	2019/1/15
Sequence similarity network	Show

Name	Gene	UniRef100	NCBI NR	Cluster	Length	Pattern match	Match score
VULLA010001	OR5C1	---	---	C01	320	🟩	★★★★★
VULLA010008	OR2AK5C	---	---	C01	316	🟩🟥🟩	★★★★★
VULLA010009	OR4C102	UniRef100_A0A3Q7UN18	XP_025868675.1	C01	314	🟩	★★★★★
VULLA010010	OR9G4D	UniRef100_F1Q197	XP_005631777.1	C01	312	🟩	★★★★★
VULLA010011	OR9G3B	---	---	C01	311	🟩	★★★★★

● Similar Protein
○ Similar Nucleotide

100% Identity 90% Identity 80% Identity

Similar ORs (356)

□	Site mutation	Sequence id	Taxonomy ID	Scientific Name
□	show mutation	TRASP010169	69298	Tragelaphus spekii
□	show mutation	TRAST010006	9946	Tragelaphus strepsiceros
□	show mutation	LAMGL010476	273914	Lama glama

Style: cartoon
Color: Atom index
ColorScale: BWR
Background: ● White ○ Black
Opacity: 1
clipNear: 0
Filter syntax: Selection Language
Backtrack

Protein sequence

● Chemistry ○ Charge ○ Hydrophobicity ○ Clear

```
    MERGNHTVTE FILVGFTTDP MMQLVLFVVF LGVYSLTVVA NATLIVLICN
51  DSRLHTPMYF FIGNLSFLDV WYSSVYTPKI LMTCISEDKS ISFAGCVCQF
101 FFSAGLAYSE CYLLAAMAYD RYAAISNPLL YAQTMSSRLC ICLVVYSYTG
151 GFVNAIILTS NTFTLDFCGD NVIDDFFCDV PPLVKLACDV KESYQAVLYF
201 LLASNVITPT ALILASYMFI IAAILRIRST QGRRKAFSTC SSHLVSVTLY
251 YGSILYIYSR PSSSYSLERD KMVSTFYTVL FPMLNPMIYS LRNKDVKEAL
301 KKLIVLTQSD V
```

Nucleotide sequence

```
1   ATG GAG AGA GGC AAT  CAC ACA GTG ACT GAA  TTC ATC CTG GTG GGC  TTC ACG ACA GAC CCA  ATG ATG CAG CTG GTC
26  CTG TTT GTG GTA TTC  CTT GGC GTG TAC TCT  CTG ACT GTG GTA GCA  AAT GCC ACC CTC ATA  GTG TTG ATC TGT AAT
51  GAC TCC CGG CTG CAC  ACA CCC ATG TAT TTT  TTC ATT GGG AAT CTA  TCT TTT CTG GAT GTC  TGG TAT TCC TCC GTC
76  TAC ACA CCA AAG ATC  CTC ATG ACC TGC ATC  TCT GAA GAC AAA AGC  ATC TCC TTT GCT GGC  TGT GTA TGT CAG TTC
101 TTC TTC TCT GCT GGG  TTG GCA TAT AGT GAG  TGT TAC CTG TTG GCC  GCC ATG GCT TAT GAC  CGC TAT GCA GCC ATT
126 TCC AAC CCC CTG CTG  TAT GCT CAG ACC ATG  TCA AGT AGA CTG TGC  ATC TGT TTG GTT GTA  TAC TCC TAT ACT GGA
151 GGT TTT GTC AAT GCA  ATA ATA CTT ACC AGT  AAC ACA TTC ACA TTG  GAT TTT TGT GGT GAC  AAT GTC ATA GAC GAT
176 TTT TTC TGT GAT GTC  CCA CCA CTG GTG AAA  TTG GCA TGT GAT GTG  AAG GAA AGC TAT CAG  GCT GTG TTG TAC TTC
201 CTC CTG GCC TCC AAT  GTC ATC ACT CCC ACG  GCG CTC ATC CTT GCT  TCC TAC ATG TTC ATC  ATT GCT GCC ATC TTG
226 AGG ATC CGC TCC ACC  CAG GGC CGT CGC AAG  GCC TTC TCC ACA TGC  TCC TCC CAC CTG GTC  TCA GTC ACT TTG TAC
251 TAT GGC TCC ATT CTC  TAC ATC TAC TCT CGC  CCA AGT TCC AGC TAT  TCC CTT GAG AGG GAC  AAA ATG GTA TCA ACA
276 TTT TAT ACT GTG CTA  TTC CCC ATG TTG AAT  CCC ATG ATC TAT AGT  CTA AGG AAT AAG GAT  GTG AAA GAA GCT CTG
301 AAA AAG CTC ATC TGG  TTA ACA CAA TCA GAT  GTC TAA
```

Figure 3 The "Receptors" menu has four levels. A, "All species in a clade" is the first level with each row giving information of one species in an evolutionary clade. Clicking the button at the end of the row enters the second level page. B, "Single species" is the second level page, which is a species detail card. C, The sequence similarity network of ORs in a species. D, The evolutionary tree of ORs in the species. E, The "All ORs in a species" is the third level page, with each row providing information of a single OR. Clicking the button at the end of the row enters the fourth level page, "single OR". F, Similar DNA or protein sequences of an OR. G, Example snake plot of an OR. H, The predicted structural model of an OR. I, The protein and DNA sequences of an OR.

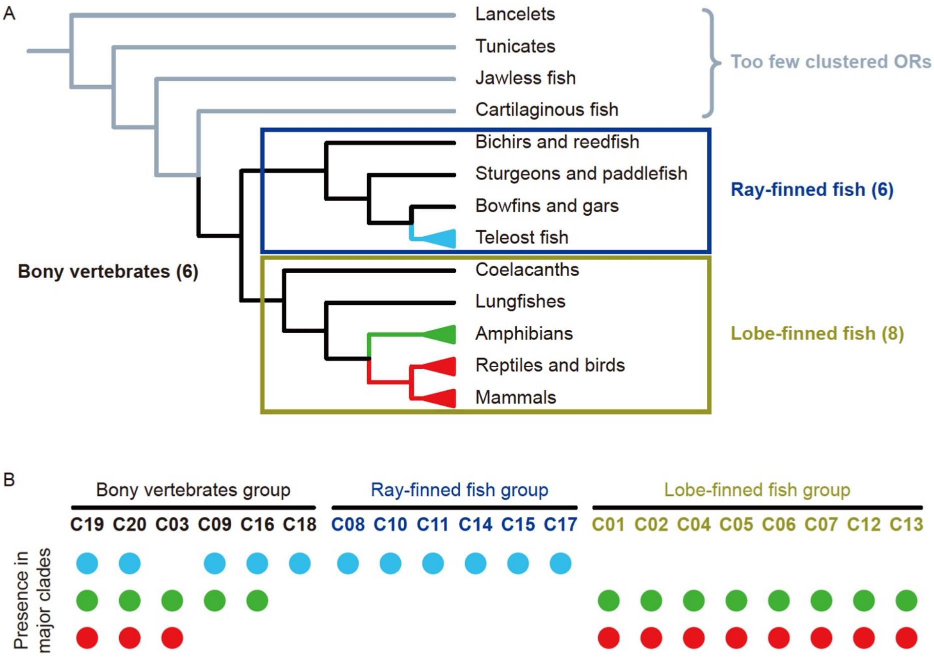


Figure 4 The distribution of OR communities in major evolutionary clades. A, The evolutionary clades of whole genome sequenced chordates. The clades with color triangles have abundant whole genome sequenced species. The clades in the blue and yellow boxes belong to "ray-finned fish" and "lobe-finned fish". B, The distribution of 20 OR communities in major evolutionary clades. The sky blue, green and red circles represent the "teleost fish", "amphibians" and "reptiles, birds and mammals" clades, respectively.

TM7 are quite conserved (corresponding to orange-coloured circles in Figure S8 in Supporting Information), inferring their roles in OR activation.

The key residues involved in non-olfactory class A GPCRs common activation pathway were previously described in details (Zhou et al., 2019). Here, we compared the similarities and differences in activation pathways between ORs and non-olfactory class A GPCRs. We note that the TM6 of ORs differ significantly from that of the non-olfactory class A GPCRs, and a comparison of their TM6 structures reveals that the registry of the TM6 of ORs is shifted by three residues towards the intracellular direction (Figure S9 in Supporting Information). In addition, it is not difficult to find that ORs do not have the conserved toggle switch residue $W^{6.48}$, instead several much smaller residues tend to appear in this position (6.51 in OR). This seems to imply that the initiation step of receptor activation in ORs is completely different from that in non-olfactory class A GPCRs.

To further explore the differences of activation mechanisms between ORs and non-olfactory class A GPCRs, we carefully examined whether the six key residue clusters in the common activation pathway of non-olfactory class A GPCRs change in ORs (Figure 5). We chose the AlphaFold2 model of a typical OR (OR5AN1), and compared it with the structure of β2AR (PDB: 2RH1), a prototype non-olfactory class A GPCR for demonstration (Figure 5). The residues forming the ligand-binding pocket of ORs are highly diverse (denoted by black triangle-marked circles in Figure S10 in Supporting Information), which reflects the diversity of OR ligands that are similar to non-olfactory class A GPCRs (Katritch et al., 2012; Ngo et al., 2017; Venkatakrishnan et al., 2013). Among ligand-binding residues, the 45.51, 6.55, and 7.41 are community conserved sites, but not globally conserved sites (Figure S10 and Data S3 in Supporting Information), inferring their potential important roles in recognizing the unique chemical groups of ligands of each OR communities. $CW^{6.48}xP$ and $P^{5.50}IF^{6.44}$ motifs are responsible for the signal initiation in non-olfactory class A GPCRs (Zhou et al., 2019). While in ORs, residues at the same positions are quite different (Figure 5A). For example, the corresponding position of the toggle switch $W^{6.48}$ in non-olfactory class A GPCRs are 6.51 in ORs, residues at this position are non-conserved smaller residues, which seem cannot play the role as a master sensor of various ligands and function as toggle switch anymore (Figure 5A). The only conserved bulky residue with 8 Å (distance between Cβ) of position 6.51 in the ligand binding pocket of ORs is $Y^{7.41}$ (Figure S11 in Supporting Information). We speculate that $Y^{7.41}$ may play the similar role in receptor activation as the toggle switch $W^{6.48}$ in non-olfactory class A GPCRs.

In non-olfactory class A GPCRs, four residues, $D^{2.50}$, $S^{3.39}$, $N^{7.45}$, and $N^{7.49}$, can directly or indirectly coordinate with Na^+ ion, forming the Na^+ binding pocket; while in ORs, residues in two out of the four positions changed: specifically, the position 3.39 is mutated to a bulkier and charged residue glutamate or aspartate, and the position 7.45 now is

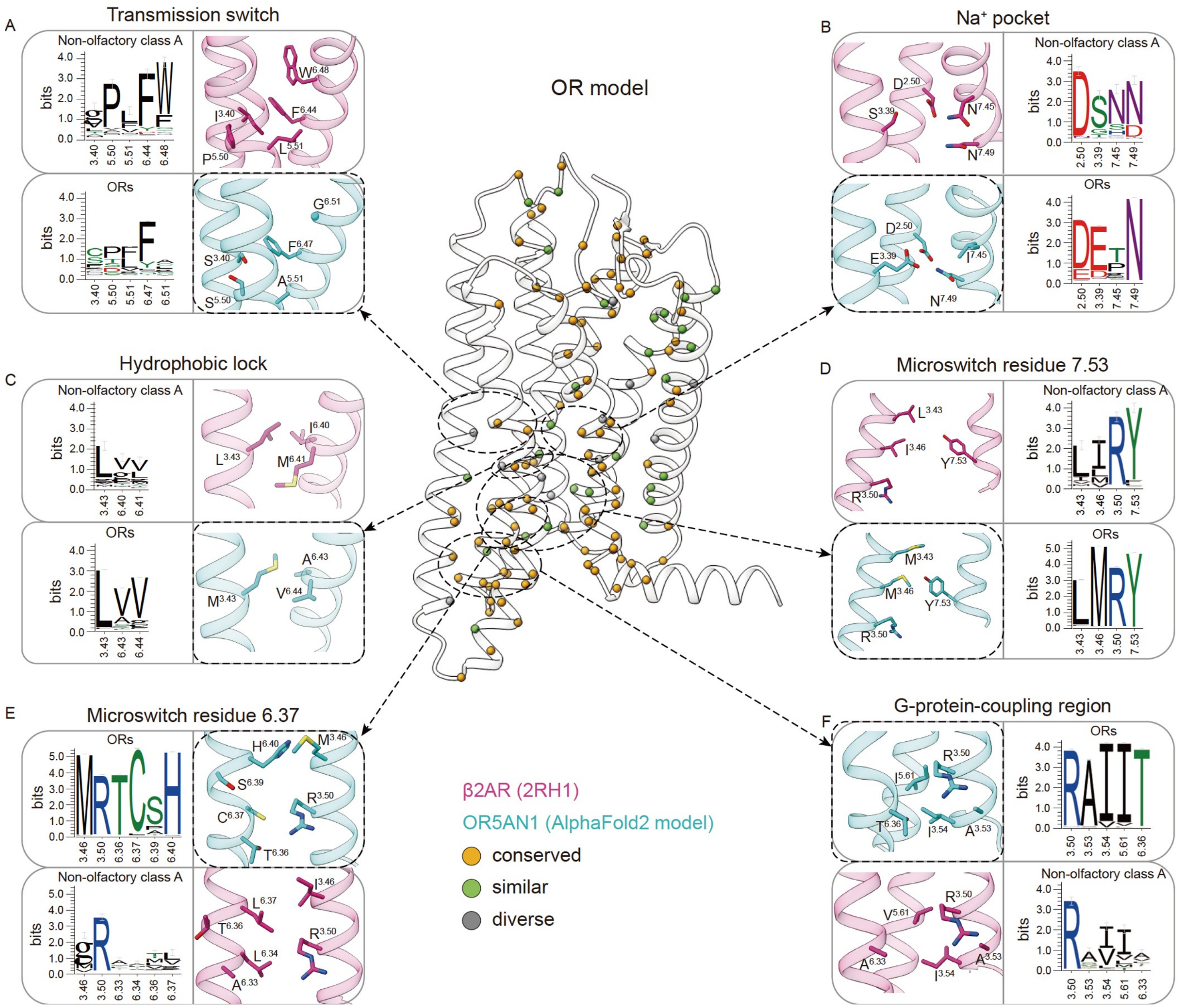


Figure 5 Comparison of sequence features between ORs and non-olfactory class A GPCRs at key positions involved in non-olfactory GPCR activation. The OR5AN1 model was retrieved from AlphaFold Protein Structure Database. Sequence and structure comparison of six key regions in receptor activation, including "transmission switch" (A), "Na$^+$ pocket" (B), "hydrophobic lock" (C), "microswitch residue 7.53" (D), "microswitch residue 6.37" (E) and "G protein-coupling region" (F).

non-conserved, dominated by threonine, proline, and isoleucine (Figure 5B). Clearly, the metal binding properties of residues in 3.39 and 7.45 dramatically changed, which implies that ORs may bind a different metal ion at this position (Figure 5B).

We have seen that the toggle switch region and the Na$^+$ binding pocket of ORs are quite different from those of the non-olfactory class A GPCRs, which tells that ORs should have unique signal initiation mechanism. Yet, residues responsible for signal amplification and propagation (layers 2 and 3 in the common activation pathway of non-olfactory class A GPCRs, (Zhou et al., 2019)) are strikingly similar (Figure 5C and D). Residues in the hydrophobic lock (3.43, 6.40, and 6.41 in non-olfactory class A, and 3.43, 6.43, and 6.44 in ORs) are still mainly hydrophobic (Figure 5C), while the microswitch residue Y$^{7.53}$ and the N$^{7.49}$P$^{7.50}$xxY$^{7.53}$ motif are the same in ORs and non-olfactory class A GPCRs (Figure 5D; Data S3 in Supporting Information).

ORs have a remarkably conserved TM6 in the intracellular half (Figure 4E; Figure S8 and Data S3 in Supporting Information), featuring H$^{6.40}$ as the most conserved site in TM6. H$^{6.40}$ is at the same position of 6.37 in non-olfactory class A GPCRs in structural alignments (Figure S9 in Supporting Information), while 6.37 is a famous microswitch residue in class A GPCRs (Venkatakrishnan et al., 2016; Zhou et al., 2019), which implies that H$^{6.40}$ may also play a key role in OR activation.

ORs recruit the same the downstream G protein, G_{olf}, which has about 76%–79% protein sequence identity with G_s. In contrast, many non-olfactory class A GPCRs have G protein selectivity, and together they can recruit various G proteins. Accordingly, five G protein-contacting positions (3.50, 3.53, 3.54, 5.61, and 6.36), are much more conserved in ORs (Zhou et al., 2019); while in non-olfactory class A GPCRs, these positions are relatively diverse (Figure 5F).

A common desensitization mechanism of GPCRs is coupling to arrestins. Non-olfactory GPCRs require phosphorylation in the C-terminus or intracellular loops (ICLs) to recruit arrestins, but most ORs do not have such phosphorylation sites primarily because they have very short C-terminals and ICLs that are not rich of phosphorylation sites Ser/Thr/Tyr. To make this clearer, we used the GPS5.0 tool (Wang et al., 2020) to perform phosphorylation site prediction for four loops (ICL1, ICL2, ICL3, and the C-terminus) in all ORs in the CORD. The predicted results show that, for most ORs, their C-terminals contain 0–2 phosphorylation sites, ICL3s and ICL1s contain 1–2 phosphorylation sites, and ICL2s contains 2–3 phosphorylation sites (Figure S12 in Supporting Information). However, from the snake diagrams (Isberg et al., 2016) and structural models of ORs, it is clear that ICL1, ICL2, and ICL3 are not long enough to dip into and bind with the positively charged groove of arrestin, thus phosphorylation sites on these loops seem not able to help recruit arrestin (Figures S13 and S14 in Supporting Information). Similarly, the OR C-terminals are either not long enough or lack of phosphorylation sites. In summary, we concluded that, unlike other GPCRs, ORs do not have proper phosphorylation sites in ICLs and C-terminals that can help recruit arrestin. ORs may have different desensitization mechanism compared with non-olfactory class A GPCRs.

DISCUSSION

Most OR functional genes only exist in one evolutionary clade

For functional ORs with gene names assigned, we carefully checked the presence of these genes in five evolutionary clades: jawed fish, amphibians, reptiles, birds, and mammals. We found that for almost all genes, they only exist in one of the five evolutionary clades. There are only two exceptions: OR5AR1 and OR6M9 exist in both reptiles and mammals.

Since about half functional ORs do not have gene names, we further checked the presence of each OR and its close homologs (no less than 80% sequence identity, if any, which means they should share a gene name if they have one) in the five evolutionary clades. The results show that ~99% ORs (386,799 out of 390,562) and its close homologs exist in only one of the five evolutionary clade, and the rest ~1% (3,763 out of 390,562) ORs and its close homologs exist in two evolutionary clades: reptiles and birds, reptiles and mammals, or birds and mammals. No ORs and its close homologs exist in more than two evolutionary clades.

The above results show that functional *Olfrs* are highly “localized”, their birth and death are usually within a single evolutionary clade, especially for those *Olfrs* in jawed fish and amphibians. For functional *Olfrs* in reptiles, birds and mammals, the dominant part of them exist only in one evolutionary clade, leaving a small portion that can exist in two evolutionary clades.

For each gene, we also counted how many species harbor this gene (Table S4 in Supporting Information). For gene names that appear in a large number of species occupying diverse ecological niches, we speculate that these OR genes are more likely ectopically expressed and play a role beyond olfaction.

Massive OR annotations and community assignments will help decipher the functions of ORs

The complex orthology relationships of ORs impede a unified nomenclature for them. Until very recently, Olender et al. (2020) introduced a unified nomenclature for ORs based on Mutual Maximum Similarity (MMS) algorithm. According to this algorithm, ORs belong to the same family share sequence no less than 40% protein sequence identity. However, this family assignment method was found to be not ideal in certain cases. For example, OR5, OR8, and OR9 should actually be one OR family in mammals and the expansion of which is related to herbivory (Hughes et al., 2018). To improve the OR family/community assignment, in this work, we divided the 404,426 functional ORs into 20 communities based on a length normalized BLOSUM80 similarity score cutoff (see MATERIALS AND METHODS). The community assignment adds an extra layer of description for OR in addition to the unified nomenclature, especially when the unified nomenclature is not available which is the case for about half functional ORs in CORD. When mapping the dozens of OR families to the 20 communities, we found that most OR families can clearly be mapped to one community (Figure S15 in Supporting Information). Of course, there are also examples that several OR families were mapped to the same OR community (Figure S15 in Supporting Information). For instance, OR5, OR8 and OR9 are all mapped to community C01, consistent with previous study that OR5, OR8 and OR9 are actually one OR family (Hughes et al., 2018). Similarly, OR1, OR3, OR7, and OR12 are all mapped to community C05, and previous study has confirmed that OR1, OR3, and OR7 should be one OR family in mammals (Hughes et al., 2018). Furthermore, OR2/10/11/13, OR41/42, OR51/52, OR61/62 are mapped to C02, C14, C03, and C10, respectively, which may imply that these communities may have also evolved with dietary niche adaption (Figure S15 in Supporting Information).

Olfr sequences are the infrastructure to study various interesting questions in the olfaction field. Here we developed a fast, accurate, and robust tool named Genome2OR that can annotate *Olfrs* from genome sequenced chordates in batch. The performance of Genome2OR is much better than that of ORA/ORFAM, in terms of accuracy and time cost. The quality of a genome assembly, especially the assembly N50, has a clear effect on corresponding annotation results. Genomes with longer assembly N50 tend to have less truncated *Olfrs*. In extreme cases, genomes have assembly N50 less than 1,000, then it is difficult for Genome2OR to identify a complete set of *Olfrs* from these genomes. Notably, RefSeq genomes generally have better quality than assembly genomes, they have no or only a few *Olfrs* annotated as truncated.

In summary, CORD provides comprehensive *Olfr* sequences for 1,695 genome sequenced chordate species. The massive *Olfr* annotations and their community assignments in this work together will help decipher functions of ORs, the majority of which still stay orphan.

MATERIALS AND METHODS

Building OR DNA profile HMMs for seven evolutionary clades

The key input for the Genome2OR tool is an appropriate profile HMM built from *Olfr* DNA sequences. We built profile HMMs for seven evolutionary clades (lancelets, jawless fish, jawed fish, amphibians, reptiles, birds, and mammals) through the following protocol (Figure S2 in Supporting Information). First, we used the profile HMM constructed from human *Olfrs* as input of Genome2OR to search against the NCBI nucleotide sequence database (NCBI nt as of February 2020) with a stringent *E*-value threshold (10^{-60}). At this threshold, it is guaranteed that all found sequences were *Olfrs*. In total, 97,281 *Olfrs* were found, and they were clustered at 80% sequence identity to generate 11,583 representative sequences. These representative sequences were then classified into the seven evolutionary clades (lancelets, jawless fish, jawed fish, amphibians, reptiles, birds, and mammals) according to their lineage information. The representative *Olfr* sequences in each clade were used to generate clade-specific multiple sequence alignment, and clade-specific DNA profile HMM (Figure S2 in Supporting Information) (Eddy, 1998).

OR annotation process

A chordate genome is sequentially passed through the three steps nhmmer.py, FindOR.py and IdentifyFunc.py to obtain the *Olfrs* of the species. Specifically, nhmmer.py uses a predefined evolutionary clade-specific DNA profile HMM to search against a user-provided chordate genome belonging to the clade to generate a hit list. FindOR.py goes through the hit list to obtain putative OR coding sequences from the genome and translates them into protein sequences. IdentifyFunc.py determines whether the putative OR coding sequences are functional or pseudo *Olfr*.

Batch.py is used to annotate multiple genomes in batch. Iteration.py provides iterative annotation with an updated profile HMM generated from previous round. Iteration.py is especially useful for non-mammalian species. Iteration.py was used for lancelets, jawless fish, jawed fish, amphibians, and reptiles in our annotation process. It is worth mentioning that these processes are encapsulated into an automated flow for ease of use.

CORD architecture

CORD is a dynamic website with frontend and backend separation. The frontend is based on Vue.js (v3.7), the backend is based on Flask (v1.1), and the database is managed by MySQL (v5.7). The frontend focuses on the layout of web pages and visualization of data, while the backend is responsible for data retrieval and processing. The separation of the frontend and backend is beneficial to the development and daily maintenance of the website, and the content of the corresponding pages will be updated synchronously when data in the database is updated, which is convenient to meet the challenge of rapid data update in the post-genome era.

Automated download of sequenced chordate genomes

We designed an automated genome download process based on the NCBI datasets tool (alpha v10.7), which can be used regularly to update the genomes used for *Olfr* annotation in the CORD. First, we used the *datasets summary* command to obtain assembly description information of all chordates (Taxon ID: 7711) in NCBI and parsed the JSON file to obtain the accessions (GCF_xxxxxxxxx or GCA_xxxxxxxxx) of any genomes that had been assembled. Then we filtered out the accessions of already downloaded genomes in our local assembly database. Finally, we used the *datasets download* and *datasets rehydrate* commands to download wanted genomes.

Genome selection for species with multiple assembled genomes

Since OR subgenomes are usually large in number and scattered in multiple chromosomes, the quality of OR annotations is deeply affected by the quality of the genome sequencing and assembly. For species with multiple available assembled genomes, we choose the genome with the best assembly quality. Specifically, if a species has a RefSeq assembly (GCF_xxxxxxxxx), then the RefSeq assembly is our first choice; if a species does not have a RefSeq assembly, then we select the assembly with the largest N50.

Assign gene names to OR DNA and protein sequences in CORD

For CORD's functional *Olfrs*, genes were named using the unified OR nomenclature developed by Doron Lancet and Tsviya Olender as adopted by VGNC (Olender et al., 2020), totally 57% OR sequences in CORD have a specific gene name. We downloaded and filtered out a total of 3,499 functional *Olfrs* from the HORDE database (Olender et al., 2013). Each of these genes possess a clear gene name. For any sequence in CORD, we calculated its sequence identity with these 3,499 genes, ignoring those results with <80% sequence identity or <95% coverage, and selected the gene with the maximum identity value from the remaining results. The gene name was then assigned to that query sequence in CORD. Thus, if none of the sequence identities was ≥80%, gene name would not be assigned.

Sequence ID mapping between CORD and UniProt/NCBI

We also mapped the CORD accession of all protein sequences to UniProt accession or NCBI accession, if applicable.

To map OR sequence IDs between CORD and UniProt/NCBI, we merged and clustered OR sequences at 100% sequence identity using the CD-HIT tool (v4.8) (Fu et al., 2012). The sequence ID mapping was resolved from the output-clustering file.

Species and lineage information in CORD

For each species in CORD, we offer its NCBI taxonomy ID, scientific name, and lineage information. The scientific name and lineage information of each species were mapped from its NCBI taxonomy ID by using the TaxonKit tool (v0.5) (Shen and Xiong, 2019).

Multiple sequence alignment for all OR protein sequences in CORD

To generate accurate multiple sequence alignment for 404,426 OR protein sequences in CORD: (i) we partitioned all functional OR protein sequences in CORD into multiple subsets, with each subset containing 300 sequences; (ii) for each subset, we chose the multiple sequence alignment file generated from all human OR proteins as the template, and made multiple sequence alignment using the "–add" and "–keeplength" parameters in MAFFT-LINSI (Katoh et al., 2002); and (iii) combined the alignment results of each subset to finalize the multiple sequence alignment of 404,426 functional OR sequences in the CORD.

Sequence similarity network

Based on the multiple sequence alignment of 404,426 functional OR sequences in the CORD, we constructed the similarity network for these sequences. In the network, each node represents an OR protein sequence, and the weight of the edge connecting any two nodes, *x* and *y*, is a length normalized similarity score S_{xy} that is calculated by

$$\mathrm{S}_{xy} = \frac{1}{L}\sum_{i=1}^{L} M(x_i, y_i),$$

where $M(x_i, y_i)$ represents the element in the BLOSUM80 matrix (Henikoff and Henikoff, 1992) for a given aligned residue pair, and L is the alignment length, here L is 310. The cutoff of similarity score for community discovery was explored between 1.5 and 2.5, with a step size of 0.1. Finally, we found that similarity score 1.8 is most appropriate.

A similar approach was also used to construct sequence similarity network for each species, and in these cases sequence identity was used as the weight of edge.

Phylogenetic tree

The phylogenetic tree of ORs for each species in the CORD was generated using MAFFT-LINSI (v7.4) (Katoh et al., 2002) and MEGACC (v10.1.8) (Kumar et al., 2018). First, multiple sequence alignment of ORs in a species was created by MAFFT-LINSI. Then gap-rich regions (if any) were filtered out. Finally, the phylogenetic tree was generated using the neighbour-joining method in MEGACC with bootstrap (1,000 replicates) statistical test.

Sequence searching

We provide protein and DNA sequence searching in CORD using the blast+ toolkit (v2.11) (Camacho et al., 2009).

Similar sequences for each functional OR sequence

Protein and DNA sequences in the CORD were clustered at 100%, 90%, and 80% sequence identities using CD-HIT tool (v4.8) (Fu et al., 2012), and the results are displayed in the CORD on the detailed description page for each receptor.

Structure modelling for all ORs in the CORD

We combined *de novo* and homology modelling approaches to construct protein structure models for the sequences in the

CORD. Our protocols were: (i) to select a few to dozens of representative sequences from each communities using CD-HIT (v4.8) (Fu et al., 2012). In total, 173 representative sequences were selected from 20 communities; (ii) *de novo* modelling of the 173 representative sequences was performed using AlphaFold2 program (Jumper et al., 2021); (iii) 48 models with <7 transmembrane helices were manually removed, and the remaining 125 models were further refined by PrepWizard in Schrodinger Suite (v2020-2) (Madhavi Sastry et al., 2013) to form the template pool; (iv) for each functional OR in the CORD, we used Modeller (v9.25) (Sali and Blundell, 1993) to build homology model for it, using the best possible template from the template pool.

Data availability

The CORD is register-free, and all sequence data can be freely browsed and downloaded at https://cord.ihuman.shanghaitech.edu.cn. Users can access the Genome2OR tool at https://github.com/ToHanwei/Genome2OR.git.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Natural Science Foundation of China (32122024, 31971178), the National Key Research and Development Programs of China (2018YFA0507000) and ShanghaiTech University. We thank Dr. Lichun Jiang and Deyu Cai from ShanghaiTech University for helpful discussions. We thank the ShanghaiTech University Library for their assistance in setting up the database.*

References

Azzouzi, N., Barloy-Hubler, F., and Galibert, F. (2014). Inventory of the cichlid olfactory receptor gene repertoires: identification of olfactory genes with more than one coding exon. BMC Genomics 15, 586.

Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65, 175–187.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res 14, 1188–1190.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics 14, 755–763.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Go, Y., and Niimura, Y. (2008). Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Mol Biol Evol 25, 1897–1907.

Gross, S.S., Do, C.B., Sirota, M., and Batzoglou, S. (2007). CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biol 8, R269.

Hayden, S., Bekaert, M., Crider, T.A., Mariani, S., Murphy, W.J., and Teeling, E.C. (2010). Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res 20, 1–9.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89, 10915–10919.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491.

Hughes, G.M., Boston, E.S.M., Finarelli, J.A., Murphy, W.J., Higgins, D. G., and Teeling, E.C. (2018). The birth and death of olfactory receptor gene families in mammalian niche adaptation. Mol Biol Evol 35, 1390–1406.

Isberg, V., Mordalski, S., Munk, C., Rataj, K., Harpsøe, K., Hauser, A.S., Vroling, B., Bojarski, A.J., Vriend, G., and Gloriam, D.E. (2016). GPCRdb: an information system for G protein-coupled receptors. Nucleic Acids Res 44, D356–D364.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059–3066.

Katritch, V., Cherezov, V., and Stevens, R.C. (2012). Diversity and modularity of G protein-coupled receptor structures. Trends Pharmacol Sci 33, 17–27.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35, 1547–1549.

Liu, A., He, F., Shen, L., Liu, R., Wang, Z., and Zhou, J. (2019). Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC Genomics 20, 977.

Liu, X., Su, X., Wang, F., Huang, Z., Wang, Q., Li, Z., Zhang, R., Wu, L., Pan, Y., Chen, Y., et al. (2011). ODORactor: a web server for deciphering olfactory coding. Bioinformatics 27, 2302–2303.

Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des 27, 221–234.

Marenco, L., Wang, R., McDougal, R., Olender, T., Twik, M., Bruford, E., Liu, X., Zhang, J., Lancet, D., Shepherd, G., et al. (2016). ORDB, HORDE, ODORactor and other on-line knowledge resources of olfactory receptor-odorant interactions. Database 2016, baw132.

Matsui, A., Go, Y., and Niimura, Y. (2010). Degeneration of olfactory receptor gene repertories in primates: no direct link to full trichromatic vision. Mol Biol Evol 27, 1192–1200.

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J.M., Irisarri, I., Wong, W.Y., Nowoshilow, S., Kneitz, S., Kawaguchi, A., et al. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. Nature 590, 284–289.

Nef, P., Hermans-Borgmeyer, I., Artières-Pin, H., Beasley, L., Dionne, V.E., and Heinemann, S.F. (1992). Spatial pattern of receptor expression in the olfactory epithelium. Proc Natl Acad Sci USA 89, 8948–8952.

Nei, M., Niimura, Y., and Nozawa, M. (2008). The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet 9, 951–963.

Ngo, T., Ilatovskiy, A.V., Stewart, A.G., Coleman, J.L.J., McRobb, F.M., Riek, R.P., Graham, R.M., Abagyan, R., Kufareva, I., and Smith, N.J. (2017). Orphan receptor ligand discovery by pickpocketing pharmacological neighbors. Nat Chem Biol 13, 235–242.

Niimura, Y. (2009a). Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Hum Genomics 4, 107–118.

Niimura, Y. (2009b). On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol 1, 34–44.

Niimura, Y. (2012). Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. Curr Genomics 13, 103–114.

Niimura, Y. (2013). Identification of olfactory receptor genes from mammalian genome sequences. In: Crasto, C., ed. Olfactory Receptors. Methods in Molecular Biology (Methods and Protocols). Totowa: Humana Press. 39–49.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res 24, 1485–1496.

Niimura, Y., and Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci USA 100, 12235–12240.

Niimura, Y., and Nei, M. (2005a). Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. Gene 346, 23–28.

Niimura, Y., and Nei, M. (2005b). Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc Natl Acad Sci USA 102, 6039–6044.

Niimura, Y., and Nei, M. (2007). Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS ONE 2, e708.

Olender, T., Jones, T.E.M., Bruford, E., and Lancet, D. (2020). A unified nomenclature for vertebrate olfactory receptors. BMC Evol Biol 20, 42.

Olender, T., Nativ, N., and Lancet, D. (2013). HORDE: comprehensive resource for olfactory receptor genomics. In: Crasto, C., ed. Olfactory Receptors. Methods in Molecular Biology (Methods and Protocols). Totowa: Humana Press. 23–38.

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

Quignon, P., Giraud, M., Rimbault, M., Lavigne, P., Tacher, S., Morin, E., Retout, E., Valin, A.S., Lindblad-Toh, K., Nicolas, J., et al. (2005). The dog and rat olfactory receptor repertoires. Genome Biol 6, R83.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118, 15.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779–815.

Salzberg, S.L. (2019). Next-generation genome annotation: we still struggle to get it right. Genome Biol 20, 92.

Sharma, V., Elghafari, A., and Hiller, M. (2016). Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. Nucleic Acids Res 44, e103.

Shen, W., and Xiong, J. (2019). TaxonKit: a cross-platform and efficient NCBI taxonomy toolkit. bioRxiv 10.1101/513523.

Skoufos, E., Healy, M.D., Singer, M.S., Nadkarni, P.M., Miller, P.L., and Shepherd, G.M. (1999). Olfactory Receptor Database: a database of the largest eukaryotic gene family. Nucleic Acids Res 27, 343–345.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32, W309–W312.

Touhara, K., and Vosshall, L.B. (2009). Sensing odorants and pheromones with chemosensory receptors. Annu Rev Physiol 71, 307–332.

Venkatakrishnan, A.J., Deupi, X., Lebon, G., Heydenreich, F.M., Flock, T., Miljus, T., Balaji, S., Bouvier, M., Veprintsev, D.B., Tate, C.G., et al. (2016). Diverse activation pathways in class A GPCRs converge near the G-protein-coupling region. Nature 536, 484–487.

Venkatakrishnan, A.J., Deupi, X., Lebon, G., Tate, C.G., Schertler, G.F., and Babu, M.M. (2013). Molecular signatures of G-protein-coupled receptors. Nature 494, 185–194.

Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., and Xue, Y. (2020). GPS 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. Genom Proteom Bioinf 18, 72–80.

Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics 29, 2487–2489.

Zhang, X. (2001). Handbook of Software Engineering and Knowledge Engineering, Vol 1. In: Chang, S.K., ed. Singapore: World Scientific Publishing Company.

Zhou, Q., Yang, D., Wu, M., Guo, Y., Guo, W., Zhong, L., Cai, X., Dai, A., Jang, W., Shakhnovich, E.I., et al. (2019). Common activation mechanism of class A GPCRs. eLife 8, e50279.

SUPPORTING INFORMATION

The supporting information is available online at https://doi.org/10.1007/s11427-021-2081-6. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.