Supporting Information

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, 393 Middle Huaxia Road, Shanghai, 201210, China

[2]School of Life Science and Technology, ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, 201210, China

[3]University of Chinese Academy of Sciences, No. 19A, Yuquan Road, Beijing, 100049, China

[4]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai, 200031, China

* Corresponding author

E-mail: zhaosw@shanghaitech.edu.cn (SZ)

Phone: +86-21-20685004

Table S1. Comparison of olfactory receptor databases. Olfactory Receptor Database (ORDB, https://senselab.med.yale.edu/ORDB), Human Olfactory Data Explorer (HORDE, https://genome.weizmann.ac.il/horde), ODORactor (http://mdl.shsmu.edu.cn/ODORactor). #func (number of functional ORs), #pseudo (number of pseudo *Olfr*), #total (number of total ORs), #species (number of species).

Database	**#func**	**#pseudo**	**#total**	**#species**
ORDB	18,735	Unknown	Unknown	70
HORDE	6,739	4,336	11,075	9
ODORactor	1,516	92	1,608	2
CORD	**404,426**	**360,822**	**765,248**	**1,695**

Table S2. Basic statistics of the CORD database.

Type	Number
Functional receptors	404,426
Functional genes	404,426
Protein models	404,426
Snake diagrams	404,426
Pseudogenes	360,820
Chordate species	1,695
Phylogenetic trees	1,695
Sequence similarity networks	1,695
Olfactory receptor communities	20

Table S3. Overview of the clades in CORD. #func (number of functional ORs), #pseudo (number of pseudo *Olfr*), #species (number of species).

Clade	#species	#func	#pseudo
Lancelets	5	36	182
Jawless fish	5	234	87
Jawed fish	647	39,350	22,058
Amphibians	19	11,862	5,228
Reptiles	63	32,956	22,459
Birds	514	27,775	35,246
Mammals	442	292,213	275,560
Total	**1,695**	**404,426**	**360,820**

Table S4. The top 20 widespread OR gene names. #species (number of species), #ORs (number of ORs).

Gene	Community	#species	Evolutionary clade	#ORs
OR51E2	C03	409	mammals	415
OR51E1	C03	405	mammals	408
OR6B1	C02	387	mammals	389
OR4E2	C04	372	mammals	372
OR52P2	C03	363	mammals	367
OR4K2	C04	359	mammals	366
OR51D1	C03	357	mammals	361
OR5AR1	C01	353	reptiles & mammals	368
OR52P1	C03	353	mammals	361
OR51R1	C03	352	mammals	355
OR10AC1	C02	351	mammals	354
OR51P1	C03	351	mammals	371
OR2K2	C02	350	mammals	351
OR10Z1	C02	349	mammals	354
OR5C1	C01	348	mammals	349
OR52B2	C03	346	mammals	353
OR2B11	C02	345	mammals	381
OR6A2	C02	344	mammals	373
OR10A4	C02	344	mammals	356
OR8B8	C01	344	mammals	350

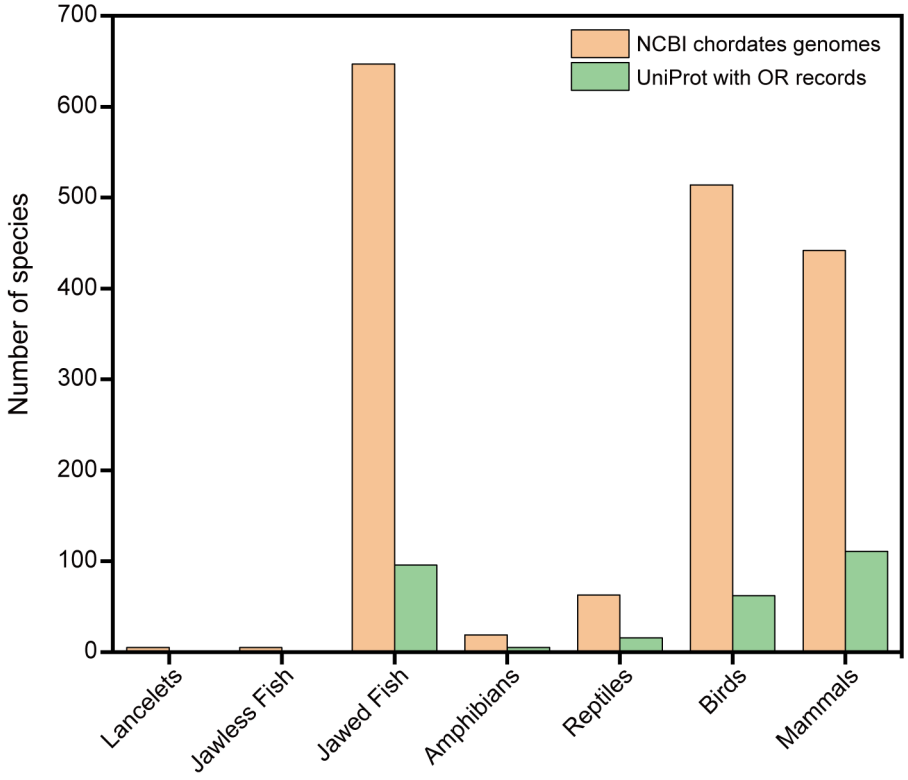


Figure S1. Comparison of the number of species with OR records in UniProt and the number of chordate genomes included in the NCBI Assembly database. (As of January 2021)

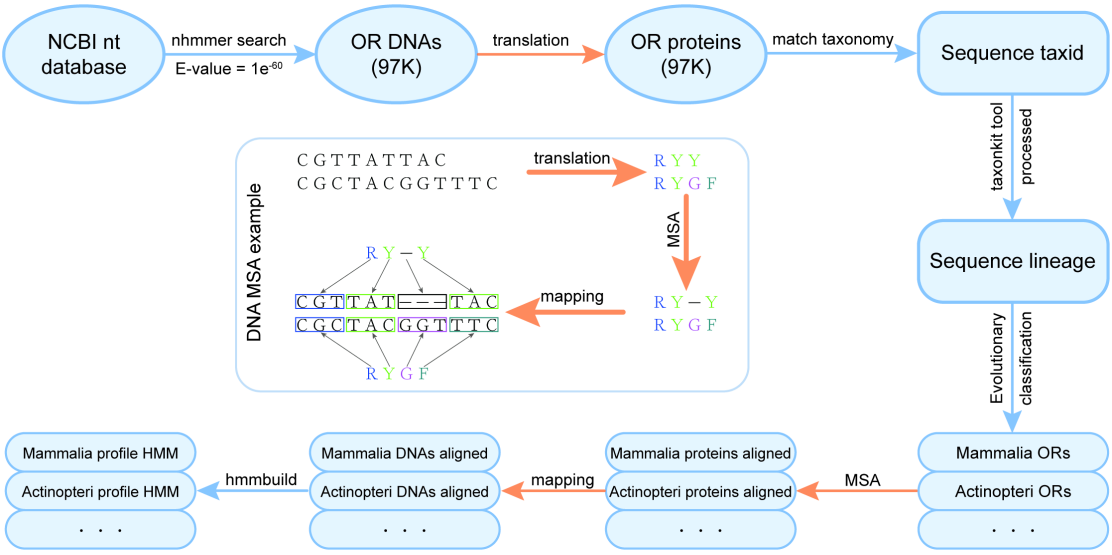


Figure S2. Steps of building a profile HMM. DNA multiple sequence alignment (MSA) was built with the guidance of protein multiple sequence alignment. DNA profile HMMs were generated by *hmmbuild* based on the MSAs.

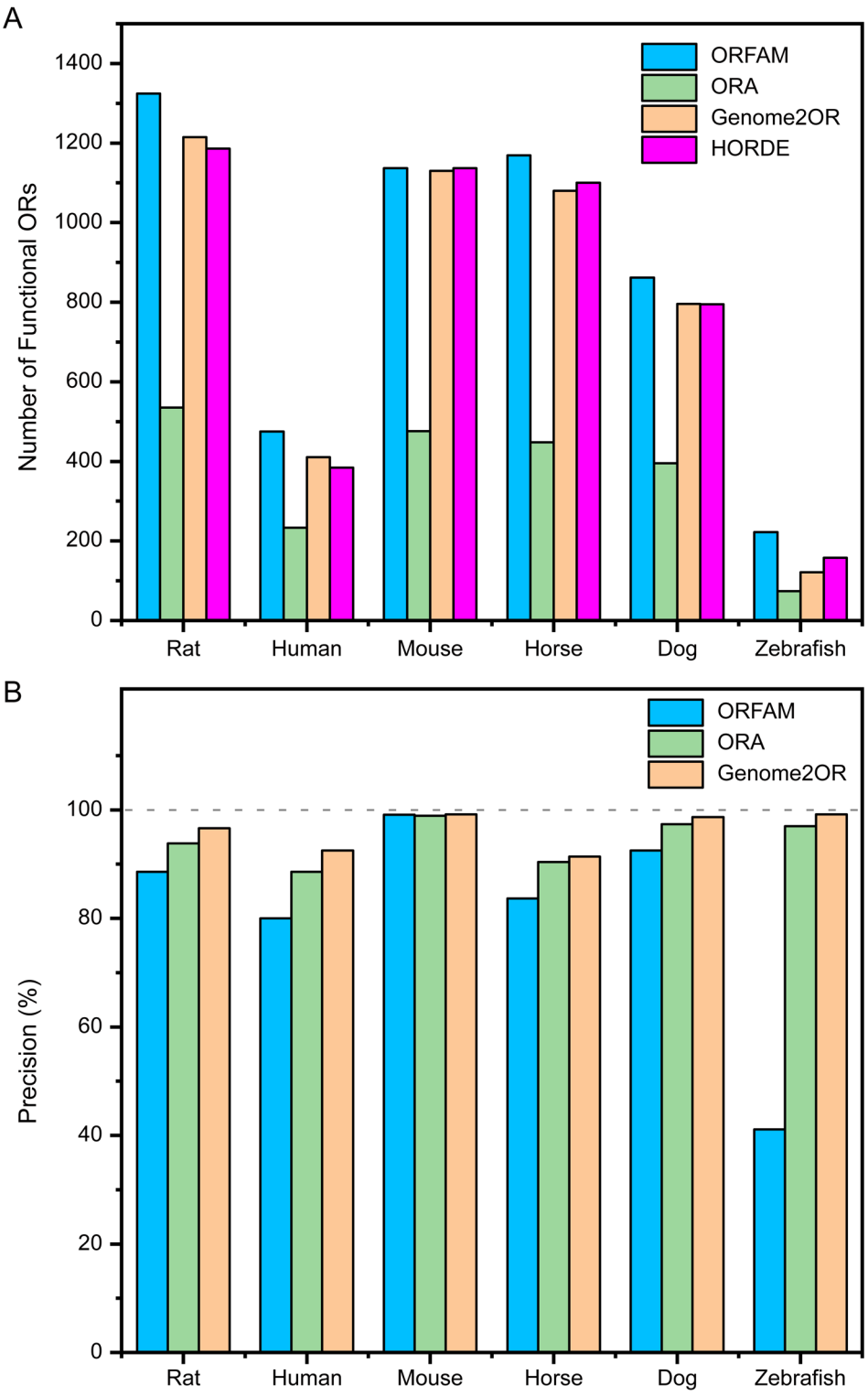


Figure S3. Comparison of Genome2OR, ORA and ORFAM annotation results. The 98% compression were used for precision comparison. See 'Supplementary Data S1' for details.

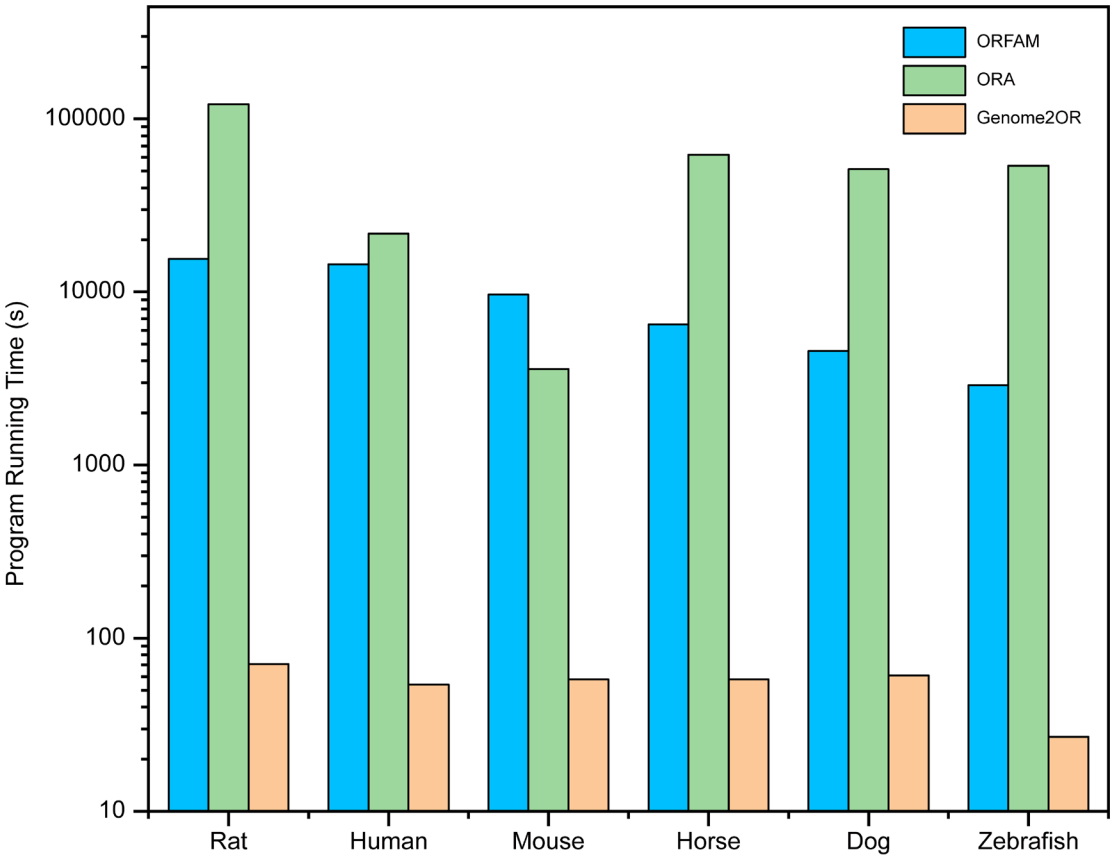


Figure S4. Comparison of time cost of Genome2OR, ORA and ORFAM. The time here is "wall clock time". Please refer to 'Supplementary Data S1' for operating environment.

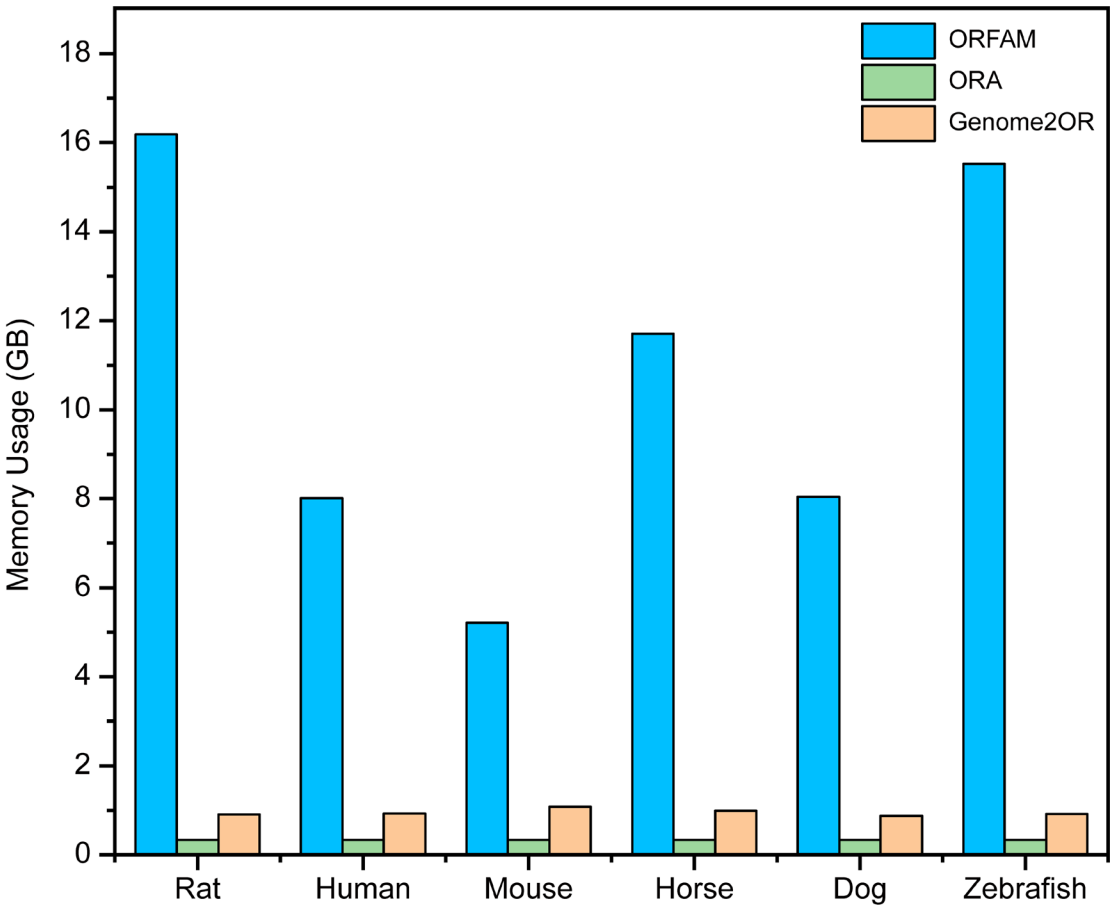


Figure S5. Comparison of memory usage of Genome2OR, ORA and ORFAM tools.

Please refer to ‘Supplementary Data S1’ for operating environment.

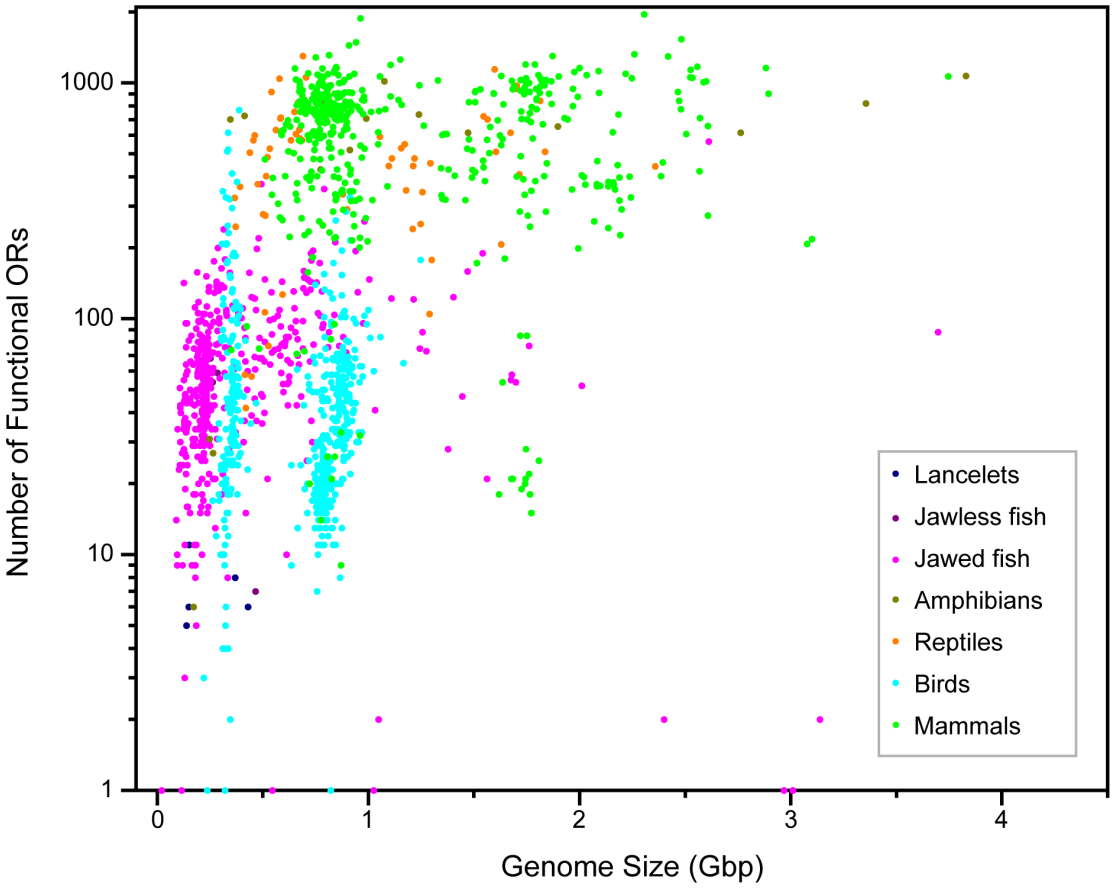


Figure S6. Relationship between the genome size of each species and the number of its functional ORs annotated by Genome2OR.

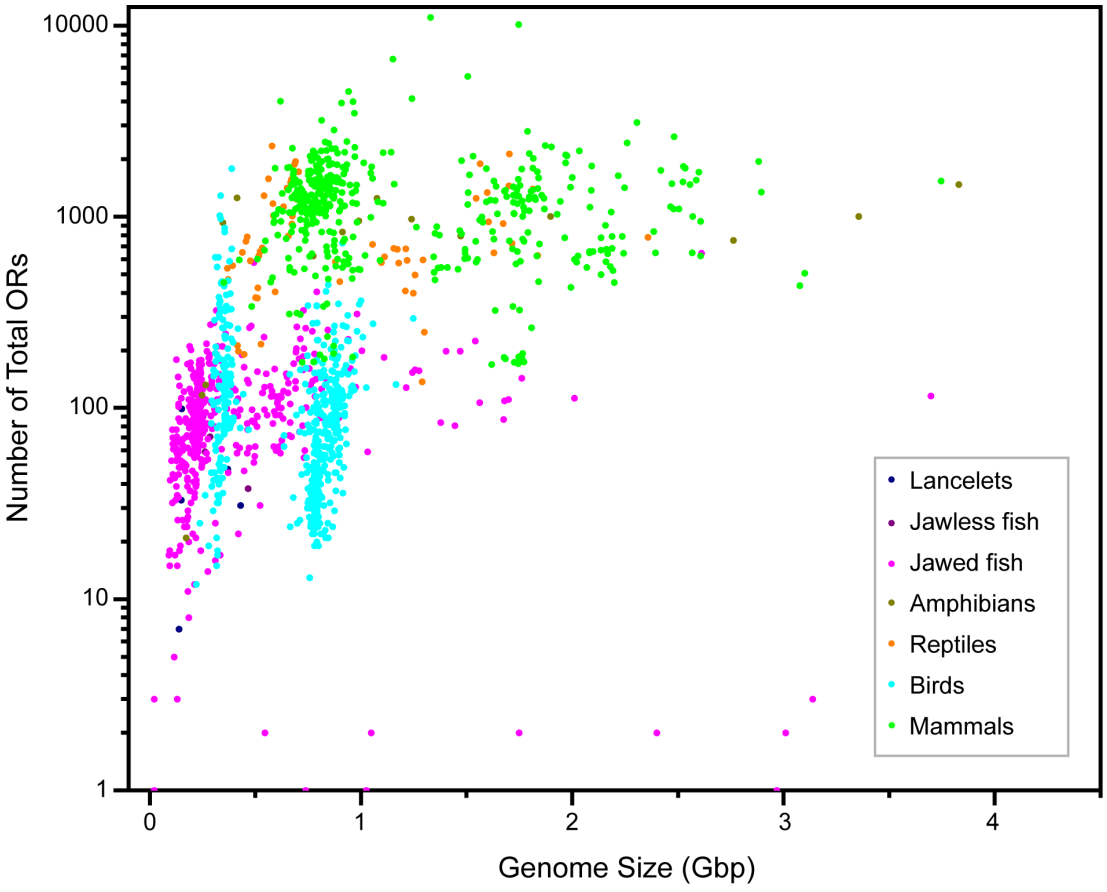


Figure S7. Relationship between the genome size of each species and the number of its total (functional and pseudogene) ORs.

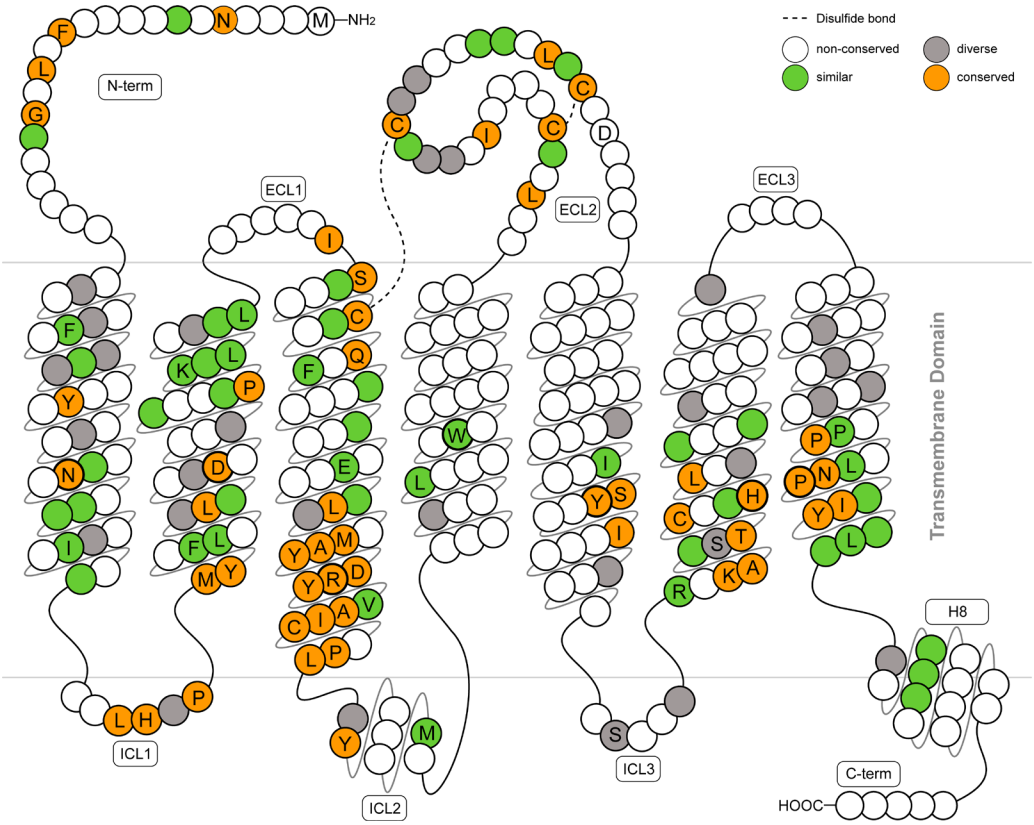


Figure S8. Snake diagram of ORs with conserved sites in colors. The residues are marked with a label indicating that more than half of the communities at this site have the same conserved residues.

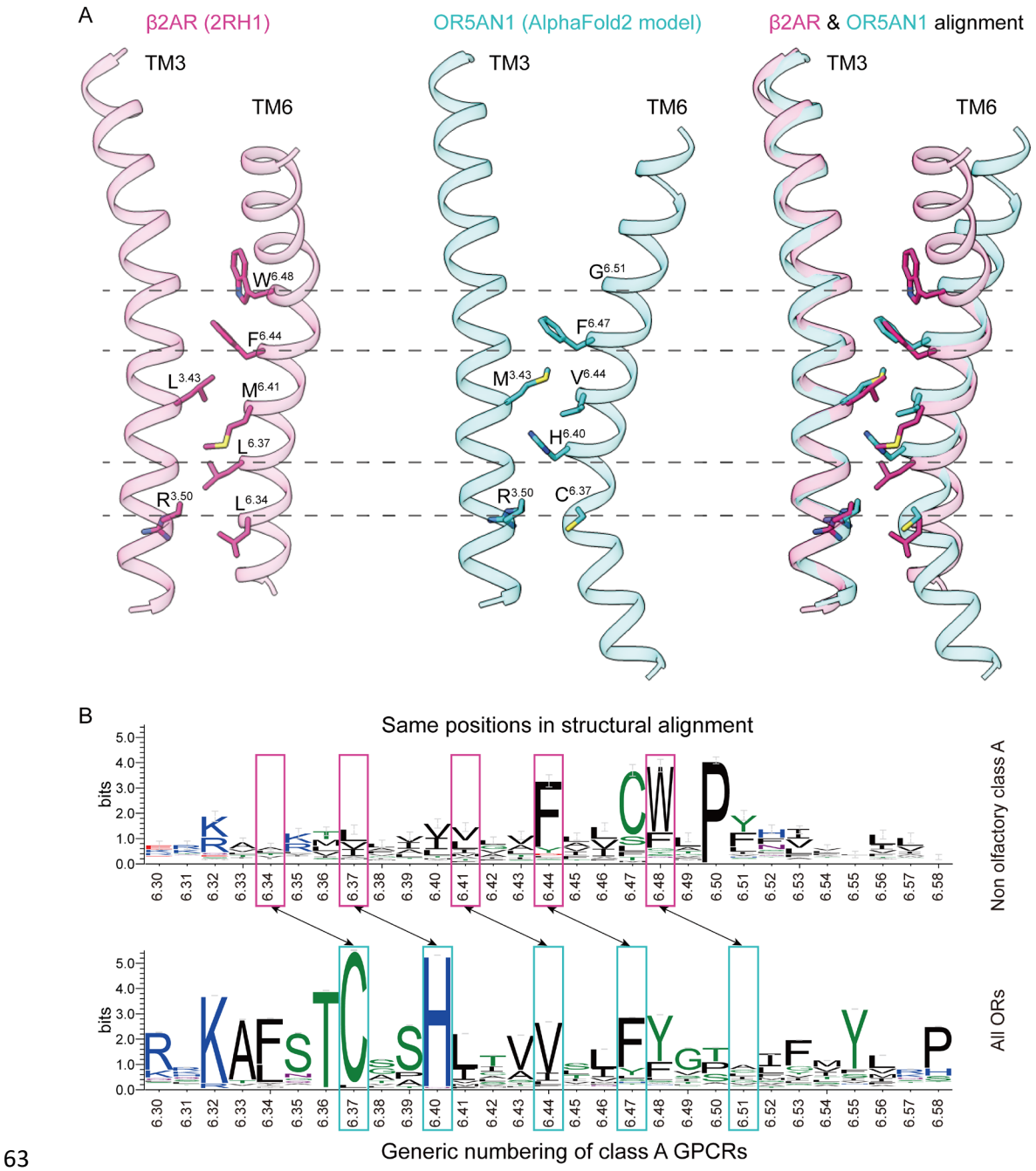


Figure S9. Structural comparison of a typical OR (OR5AN1) with a typical non-olfactory class A GPCR (β2AR). For clarity, only TM3 and TM6 are displayed. (A) Conformation and structural alignment of several activation-related residues on TM3 and TM6 for the 2RH1 and OR5AN1 models. (B) Comparison of the TM6 weblogos of non-olfactory class A and ORs.

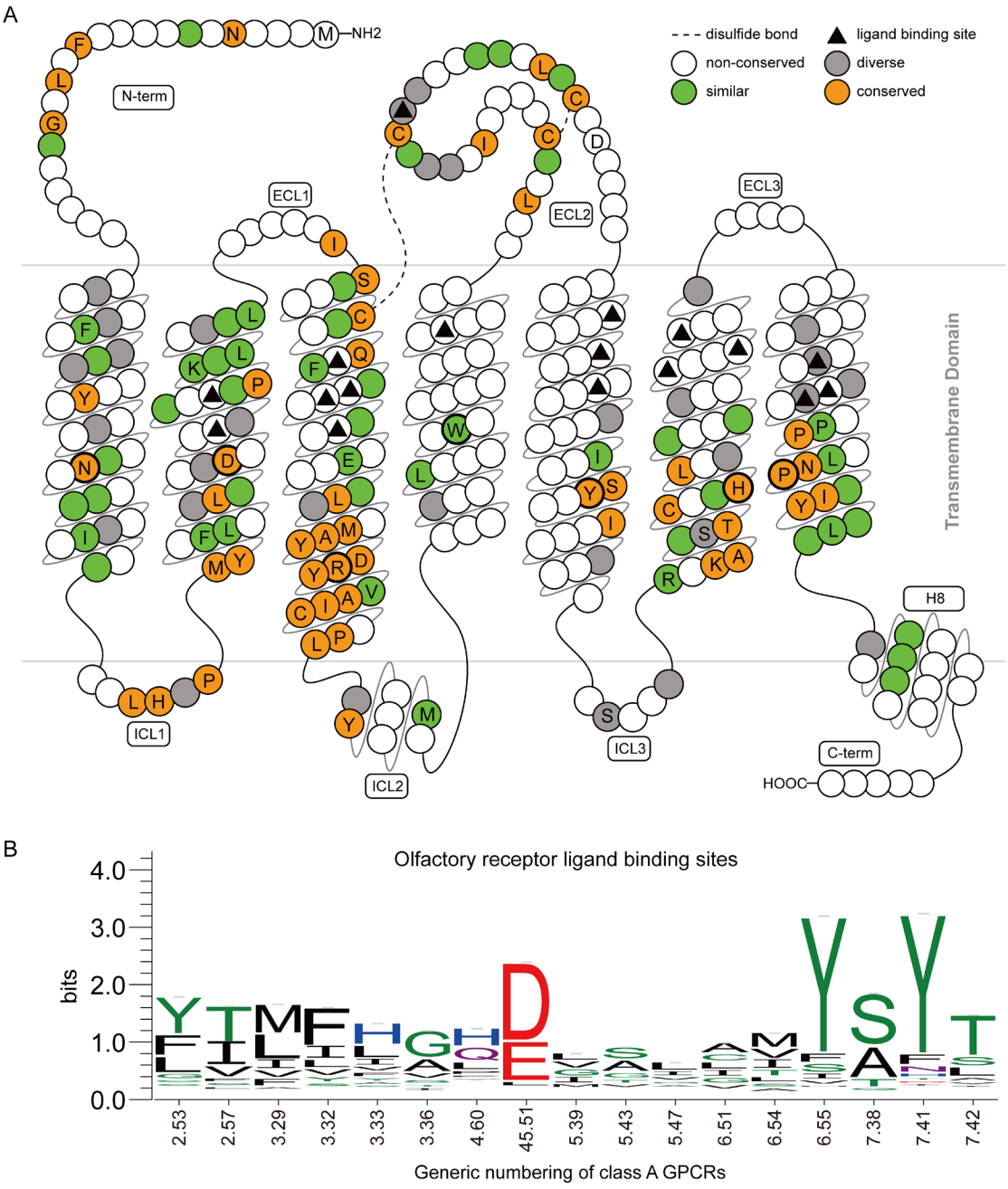
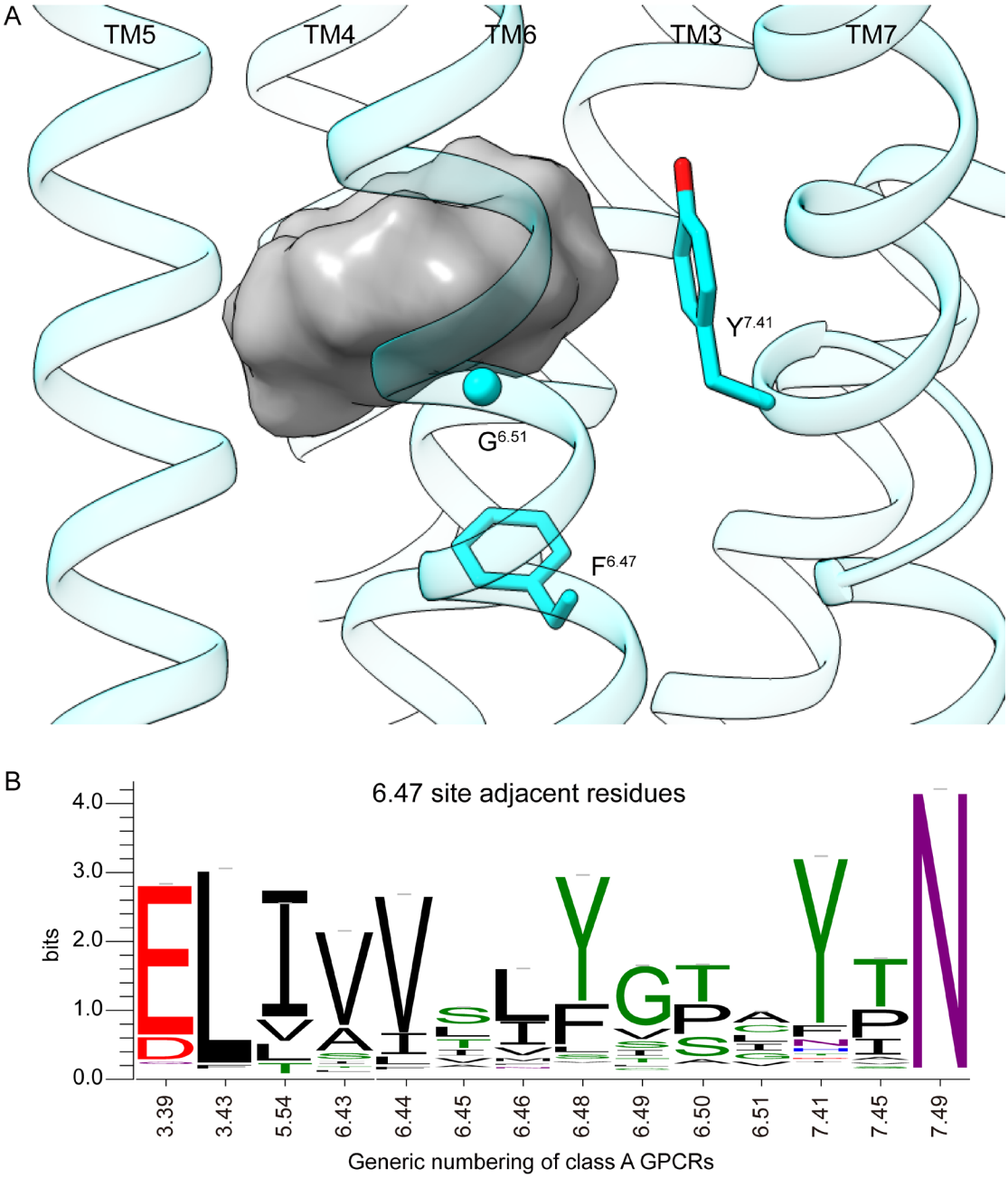


Figure S10. Ligand-binding pocket of ORs. (A) Mapping the positions of the ligand binding residues (black triangles) show that they are highly diverse. (B) The weblogo of ligand-binding pocket of ORs.

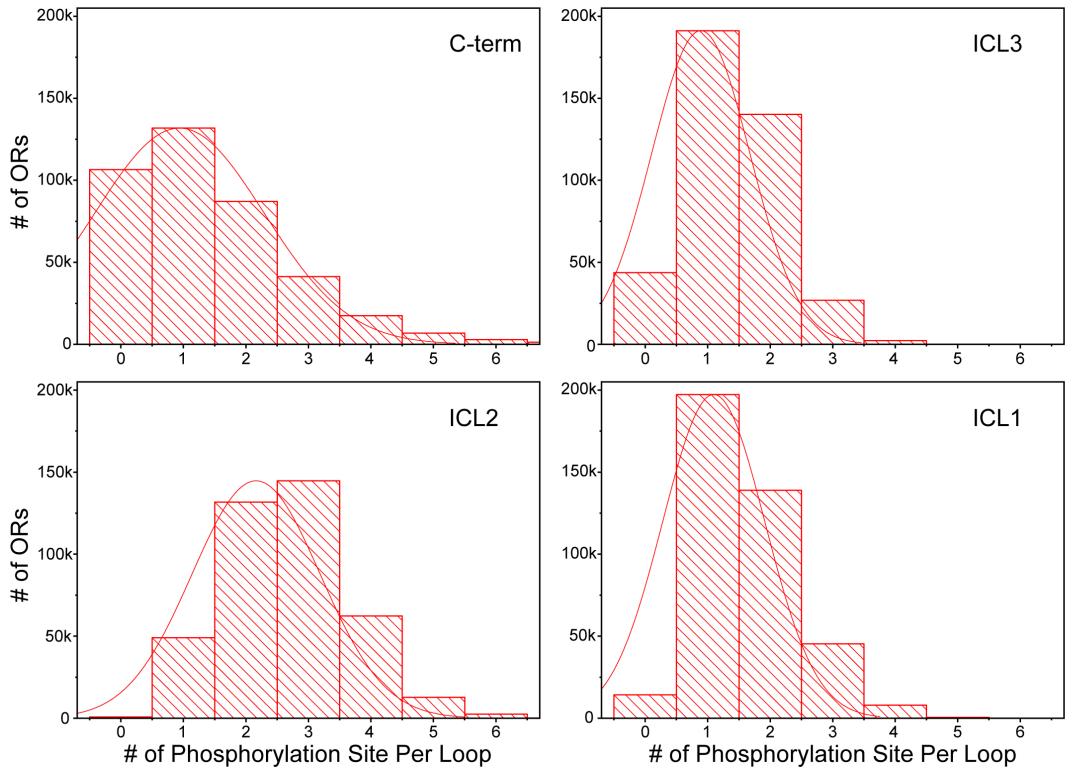


73

74 **Figure S11. Pocket facing bulky residues around 6.47 of ORs.** (A) The structural

75 relationship between positions 6.47 and 7.41. (B) The weblogo of residues around 6.47

76 of ORs (distance between Cβ is less than 8 Å).

77 

78 **Figure S12. Distribution of the number of phosphorylation sites in ICL1, ICL2,**

79 **ICL3, and the C-terminal in 404,426 olfactory receptors in CORD.**

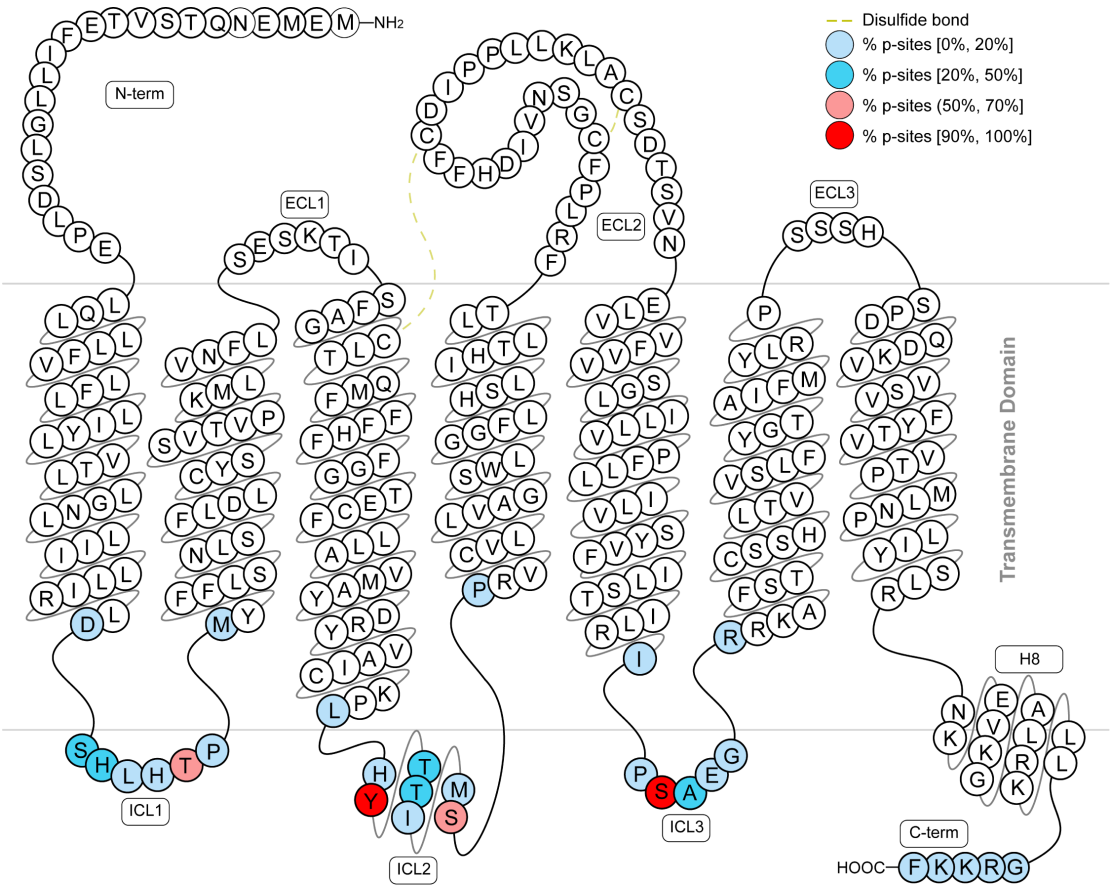
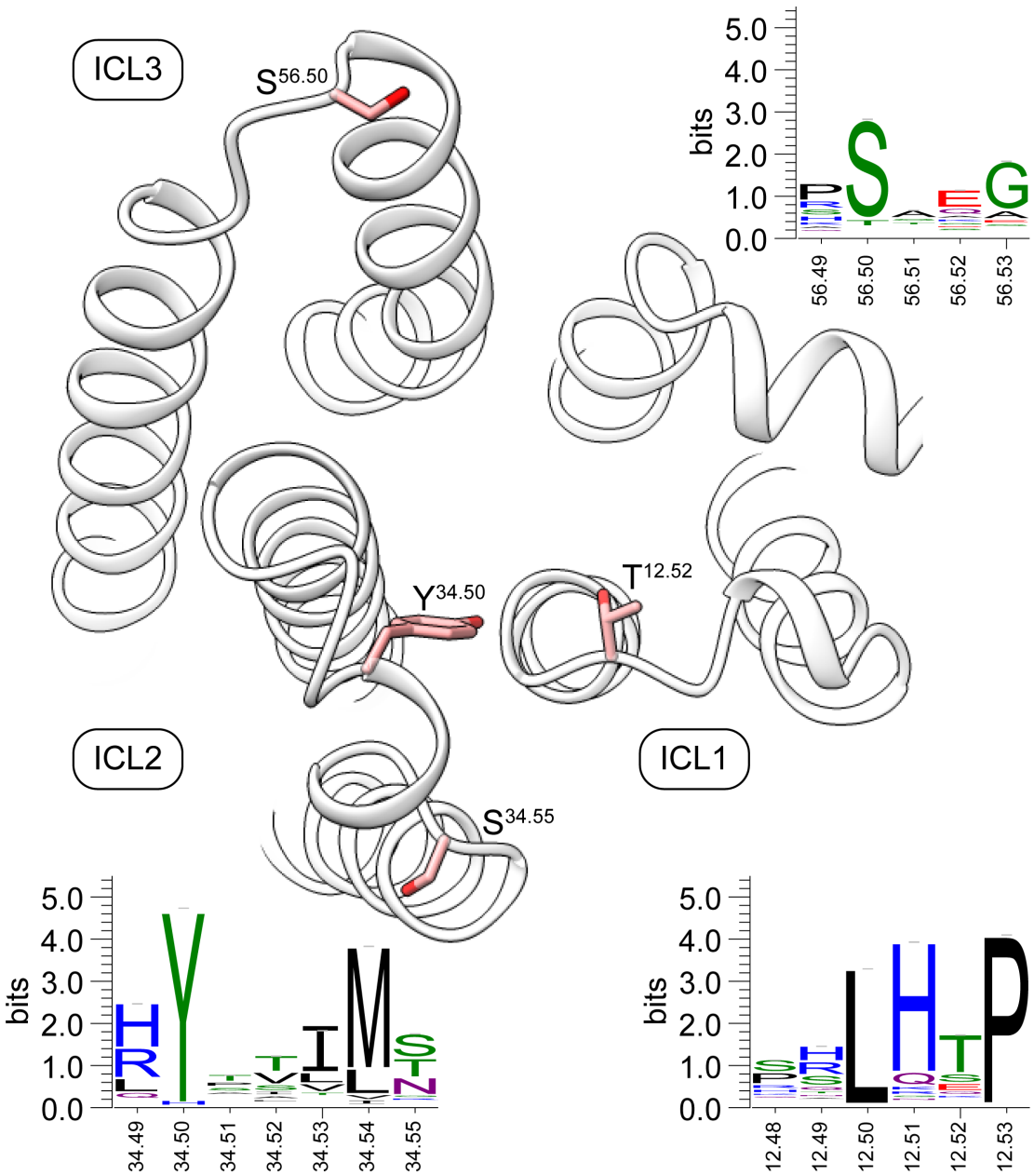
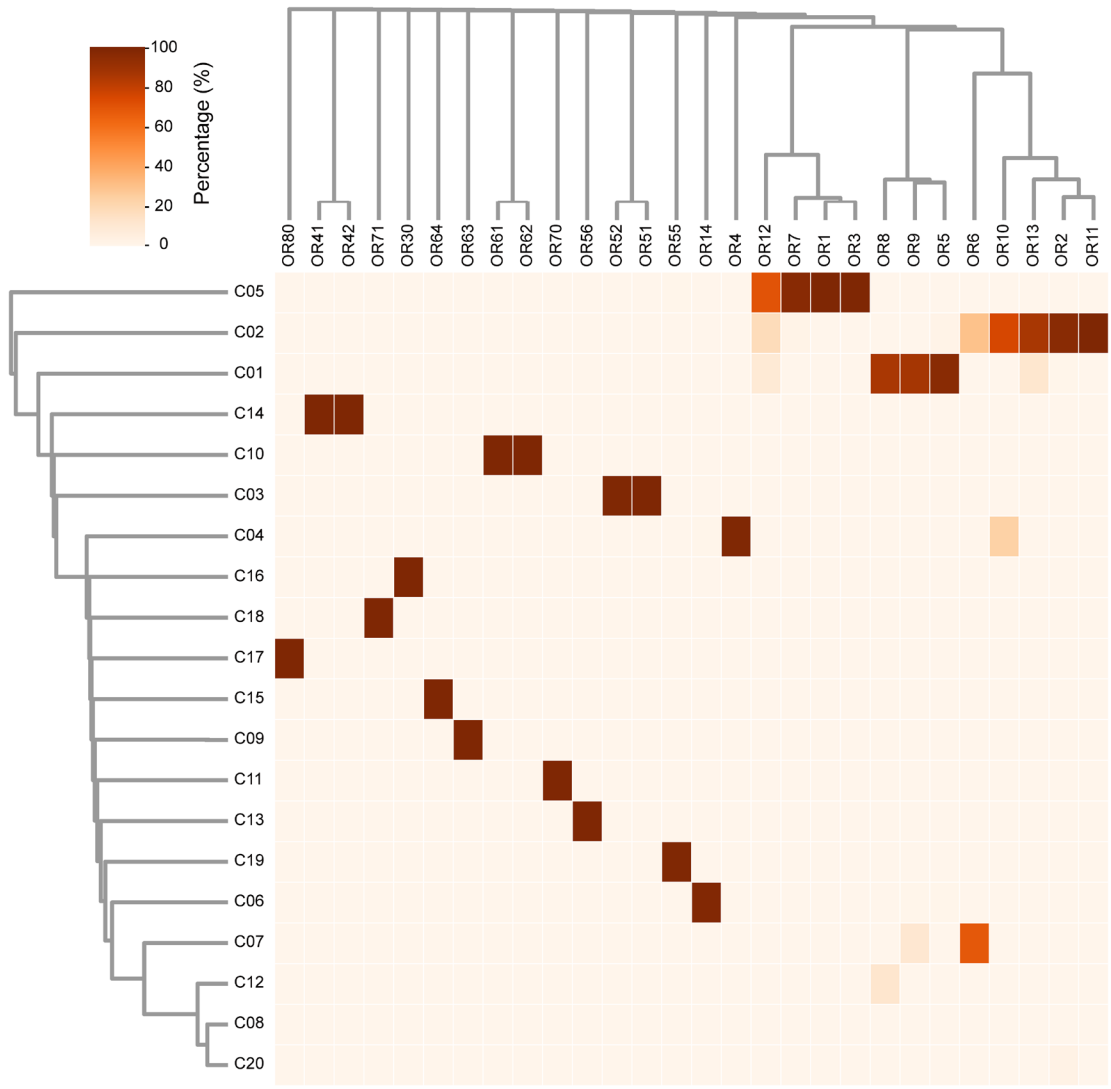


Figure S13. Frequency of occurrence of phosphorylation sites at each site of ICL1, ICL2, ILC3, and C-terminus of all ORs in CORD, mapping to the consensus snake diagram.



84

85 **Figure S14. The potentil phosphorylation sites on the OR5AN1 model.**



86

87 **Figure S15.** Mapping OR families to CORD communities.

Supplementary Data S1. Comparison of results and performance of Genome2OR and ORFAM

Supplementary Data S2. AlphaFold model of ORs for each community representative receptor.

Supplementary Data S3. Comparison of ORs and weblogo of various communities with non-OR class A GPCR.

Supplementary Data S4. Table of conserved sites for each OR community of CORD.