Comprehensive, structurally-curated alignment and phylogeny of Vertebrate biogenic amine receptors

Stephanie J. Spielman[1,2,3], Ahmad R. Sedaghat[4,5], Keerthana Kumar[1,2,3], and Claus O. Wilke[1,2,3]

[1]**Department of Integrative Biology, The University of Texas at Austin, Austin, U.S.A.**
[2]**Institute of Cellular and Molecular Biology, The University of Texas at Austin, Austin, U.S.A.**
[3]**Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, U.S.A.**
[4]**Department of Otolaryngology–Head and Neck Surgery, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, U.S.A.**
[5]**Department of Otology and Laryngology, Harvard Medical School, Boston, Massachusetts, U.S.A.**

ABSTRACT

Keywords: biogenic amine receptors, phylogenetics, multiple sequence alignment, G-protein coupled receptors

INTRODUCTION

Biogenic amines, such as the molecules serotonin and dopamine, play critical roles in virtually all Metazoa taxa, exerting significant influence on both behavior and physiology. In vertebrates, these molecules' actions are mediated primarily through the biogenic amine receptor family, which includes dopamine (DRD), histamine (HRH), trace (TAAR), adrenergic (ADR), muscarinic cholinergic (mAChR), and most serotonin receptors (5HTR). Biogenic amine receptors belong to the broad family of G protein-coupled receptors (GPCRs), one of the largest and most diverse receptor families in eukaryota, and indeed Metazoa. Indeed, due to the extensive diversity of biological functions they direct and the ongoing expansion of their ligand repertoire, GPCRs are considered one of the most evolutionarily innovative and successful gene families (Bockaert and Pin, 1999; Lagerstrom and Schioth, 2008).

Biogenic amine receptors form a clade within large the Rhodopsin-like (family "A") GPCR family, whose emergence likely accompanied that of the Opisthokont (Fungi and Metazoa) lineage (Krishnan et al., 2012). Subsequently, the Rhodopsin-like family underwent substantial expansion in Metazoa, most notably in Vertebrate lineages (Rompler et al., 2007; Stäubert et al., 2013). Like all GPCRs, these receptors contain a highly-conserved, characteristic GPCR structure of seven transmembrane (TM) domains separated by three extracellular (ECL) and three intracellular (ICL) loops, and they propagate intracellular signaling through a G protein-mediated pathway. Biogenic amine receptors are prominent targets for a wide array of pharmaceuticals aimed to treat myriad diseases such as schizophrenia, migraines, hypertension, allergies and asthma, and stomach ulcers (Schoneberg et al., 2004; Evers et al., 2005; Mason et al., 2012).

In spite of these receptors' biological and clinical importance, studies on their evolution are relatively limited. Evolutionary studies which have focused on a biogenic amine receptors have predominantly been limited to single receptor subtypes, namely TAAR (Gloriam et al., 2005; Lindemann et al., 2005; Hashiguchi and Nishida, 2007), DRD (Callier et al., 2003; Yamamoto et al., 2013), and 5HTR Anbazhagan et al. (2010). Moreover, many of these studies, and indeed larger-scale studies on the general evolution of the Rhodopsin-like GPCR family, have examined very narrow species distributions, for instance specifically teleosts (Gloriam et al., 2005), primates (Anbazhagan et al., 2010), humans and mice (Vassilatis et al., 2003; Kakarala and Jamil, 2014), or even strictly humans Fredriksson et al. (2003). Thus, studies which account for the full breadth of known bioamine receptor sequences as well as their broad species distributions remain elusive.

This dearth of biogenic amine receptor-specific evolutionary understanding is underscored by the difficulties in establishing a robust multiple sequence alignment (MSA). MSAs provide the founation for nearly all comparative sequence analyses, and they are commonly used to locate conserved sequence motifs, identify functionally important residues, and investigate the evolutionary histories. As constructing an MSA represents the first step in any sequence analysis, MSA errors

are known to bias these downstream analyses (Ogden and Rosenberg, 2006; Wong et al., 2008; Jordan and Goldman, 2012). It is therefore crucial to ensure accuracy in MSAs to the extent possible. For GPCR sequences, in particular, any MSA should recapitulate the canonical 7TM structure, which a naive alignment of sequences cannot necessarily accomplish. While there are certain MSA software platforms that explicitly incorporate structural information into the alignment algorithm (e.g. 3DCoffee (O'Sullivan et al., 2004) and PROMALS3D (Pei et al., 2008)), these programs are fairly computationally-intensive and thus ill-suited for large-scale (over 1000 sequences) applications. Moreover, such programs require the use of a single crystal structure to guide sequence alignment. While all GPCRs have the same conserved 7TM domains, different GPCR subfamilies, particularly the biogenic amine receptors, feature a wide variety of ICL and ECL sizes. For instance, the ECL3 lengths for human HRH1 and DRD3, respectively, are roughly 27 and 117, and their respective ICL3 lengths are 68 and 14 (as predicted by GPCRHMM (Wistrand et al., 2006)). Thus, aligning with a single crystral structure will not effectively represent the domain variability across biogenic amine receptor subtypes. A desirable alignment strategy would instead anchor all sequences by their conserved 7TM domains without inappropriately constraining the heterogeneous ECL and ICL domains.

We therefore have adopted a novel iterative alignment strategy to create a large (3064 sequences), structurally-curated MSA of vertebrate biogenic amine receptors. In order to ensure proper structural alignment, we employed the software GPCRHMM (Wistrand et al., 2006), which uses a hidden markov model approach to assign each residue in a given GPCR sequence to its respective domain, either extracellular, transmembrane, or intracellular. Previously validated with resolved Rhodopsin-like GPCR crystal structures (Spielman and Wilke, 2013), GPCRHMM relies on the remarkably-conserved GPCR 7TM structure to predict GPCR domains from protein sequence alone with exceptional accuracy. Our alignment strategy therefore ensured that all predicted 7TM domains aligned appropriately across all sequences. Importantly, this strategy did not require any manual or visual data inspection, thus avoiding any confounding subjectivity in MSA processing. Moreover, to demonstrate the utility of our alignment, we constructed a phylogeny of our receptors with this alignment as well as a naive MSA which did not undergo this iterative process. We found that our structurally-curated MSA offered dramatic improvements in phylogenetic fit relative to a structurally-naive MSA. Furthermore, through this structurally-aware phylogeny, we are able to discern relationships among biogenic amine receptor subtypes with a far increased level of sensitivitiy relative to previous studies. We additionally identify novel lineage-specific receptor clades and clarify NCBI annotations for over 30 sequences.

We present this large, taxonomically-comprehensive vertebrate biogenic receptor MSA and its corresponding phylogeny as a resource for any group interested in studying the dynamic evolutionary processes, structural, and/or functional constraints operating within this exceptionally important GPCR clade. All data, including MSAs, phylogenies, and sequence descriptions, as well as all code used to generate these data freely available from `https://github.com/sjspielman/amine_receptors`. This structurally-aware MSA and corresponding phylogeny represent the most comprehensive and curated vertebrate biogenic amine receptor dataset to date and should therefore be extremely useful for studying both the broad patterns governing this biogenic amine receptor sequence evolution as well as receptor-specific evolutionary trends. Further, our curated MSA should may serve as a helpful resource in the ongoing development of homology models and pharmaceutical therapeutics targeting these receptors (Kristiansen, 2004; Ishiguro, 2004; Evers et al., 2005; Mason et al., 2012).

RESULTS

Constructing a structurally-curated MSA of biogenic amine receptors

We collected all sequences using PSI-BLAST and filtered data to exclude poor-quality sequences and/or sequences with excessive ambiguities to ensure a high-quality data set (see *Methods* for details). We additionally retained only sequences that could be unequivocally classified as GPCRs using the program GPCRHMM (Wistrand et al., 2006), leaving a total dataset of 3464 receptor sequences to align. We then built a structurally-curated MSA from these 3464 receptor protein sequences according to the iterative strategy outlined in Figure 1. Before aligning sequences, we again used the program GPCRHMM (Wistrand et al., 2006) to assign each residue in all protein sequences to its respective structural domain (extracellular, transmembrane, or intracellular) using a 0.5 posterior probability cutoff. Next, for each iteration of the algorithm in Figure 1, we aligned protein sequences with MAFFT (Katoh and Standley, 2013). Using the GPCRHMM-derermined residue domains, we assigned a consensus structural domain to each MSA column. We then discarded all sequences for which $\geq 5\%$ of residues did not correspond to their respective column's consensus domain. We realigned the remaining sequences with MAFFT and continued in this manner until no more sequences were discarded, resulting in a structurally-curated MSA of 3039 sequences.

Using this final MSA, we additionally created a "masked" MSA in which protein residues which did not conform to their respective consensus domains were replaced with a "?". By replacing these positions with an ambiguous character, we ensuring that each MSA column strictly contained residues belonging to the same structural domain. In total, 2.69% of all MSA positions were were masked in this second structural MSA. These final structurally-curated masked and unmasked

MSAs contained a total of 3039 sequences. Figure 2 displays the overall taxonomic distribution of sequences in these MSAs.

Structurally-aware approach strongly improves phylogenetic inference

To demonstrate the utility of analyzing biogenic amine receptors data with a structurally-curated MSA, we constructed five distinct maximum likelihood (ML) phylogenies, as detailed in Table 1, in RAxML. First, we built a phylogeny using an MSA, again built with MAFFT (Katoh and Standley, 2013), comprised of the full set of 3464 receptors; in other words, this MSA was not subjected to the iterative process shown in Figure 1. As this MSA was not structurally-curated, we refer to it as the "naive" MSA. We additionally constructed two phylogenies each from the structural unmasked and masked MSAs. Previous work has shown that both structural and functional constraints impose differing selection pressures in TM vs. extramembrane (EM), which consists of both extracellular and intracellular, domains, additionally producing distinct amino-acid frequency distributions in each domain class Tourasse and Li (2000); Stevens and Arkin (2001); Julenius and Pedersen (2006); Oberai et al. (2009); Spielman and Wilke (2013); Franzosa et al. (2013). As our structurally-curated MSAs allow us to precisely identify each MSA column as either TM or EM, we can conduct far more rigorous phylogentic inference using a partition analysis. Therefore, for each of our two structurally-curated MSAs (masked and unmasked), we inferred two ML phylogenies: one with two partitions representing TM and EM columns and one with a single partition for the entire alignment. Importantly, such a partitioned analysis would not be possible without a structurally-curated MSA, as we present here.

As assessed by AIC scores, the structurally-curated masked MSA yielded a far superior phylogeny compared to all other MSAs (Table 1), highlighting the benefits of structurally-aware analyses. That the masked structural MSA produced phylogenies with substantially better fits than did the unmasked MSA underscores that any structurally-aware study must be undertaken cautiously. While partitioning the MSA based on structural domains was clearly benefically, ensuring that each MSA column contained strictly residues belonging to the same domain was critical. Having even a few TM residues in a column assigned to the EM partition, or vice versa, strongly hindered phylogenetic fit, explaining the improved performance of the masked structural MSA.

Structurally-aware phylogeny reveals unknown biogenic amine receptor relationships and clades

Our resulting phylogeny, shown in Figure 3, represents the most comprehensive and curated vertebrate biogenic amine receptor phylogeny to date. This tree broadly captures many known features of biogenic amine receptor evolution, in particular that these receptors do not cluster based on ligand-binding but rather have undergone extensive functional convergent evolution. Indeed, our phylogeny reveals that only two ligand-based receptor classes, mAChR and TAAR, are truly monophyletic. Using this phylogeny, we were able to reclassify several misannotated sequences (Table S1) as well as uncover an entirely unknown clade of biogenic amine receptors. This unknown clade, sister to HRH2, contains strictly avian sequences as well as one *Xenopus tropicalis* sequence. Two evolutionary scenarios may explain this taxonomic distribution: either this clade emerged after the divergence of teleosts but was secondarily lost in Reptilia and Mammalia, or this clade represents an avian-specific diversification which the *Xenopus tropicalis* sequence resembles only convergently. Interestingly, the vast majority of sequences in this clade were annotated in NCBI as either octopamine or No9-like receptors, both of which are insect-specific biogenic amine receptors that do not occur in vertebrate taxa. This sequence misannotation reveals an intriguing case of convergent evolution and suggests the hypothesis that these receptors may interact with atypical ligands in vertebrate lineages.

Our phylogeny features remarkably high bootstrap support for each distinct clade of receptor subtypes. We additionally find very strong support for three deeper nodes in the phylogeny that reveal the relationships among distinct receptor subtypes. The first contains the three clades HRH1, mAChR, and HRH-3,4, the second contains the clades 5HTR-1, 5HTR-5, and 5HTR-7, and finally the third contains the 5HTR-4 and TAAR clades. Previous studies have yielded conflicting phylogenetic placements for the 5HTR-7 clade; some have argued that 5HTR-7 is phylogenetically distinct from all other 5HTR sequences (Kakarala and Jamil, 2014), while others have found that evidence for a single clade containing 5HTR-5,7 as a sister taxa to a clade containing ADRA1 sequences (Fredriksson et al., 2003). Alternatively, we find moderate-to-strong support for the 5HTR-7 clade having originated before subsequent diversification into 5HTR-5 and 5HTR-1, and we find full support showing that ADRA1 sequences form an entire distinct monophyletic group outside all other vertebrate biogenic amine receptors. In addition, as previously mentioned, our phylogeny reveals that HRH-3,4 is actually single monophyletic group. Moreover, the HRH-4 clade contains strictly mammalians sequences, including monotreme sequences, wheres the HRH-3 sequences are broadly distributed among vertebrate taxa. We therefore hypothesize that HRH-4 arose from an HRH-3 duplication concurrent with mammalian origins.

Of particular interest in our phylogeny are the unique evolutionary patterns revealed within the TAAR clade. While TAAR sequences do cluster together, the relationships among TAAR subtypes are highly dynamic, reflecting the extensive expansion and contraction events characterizing this receptor family's evolution (Lindemann et al., 2005; Hashiguchi and

Nishida, 2007; Staubert et al., 2010; Stäubert et al., 2013). Figure 4 displays the subtree containing specifically TAAR sequences and in fact differs somewhat from previously proposed TAAR trees (Lindemann et al., 2005; Hashiguchi and Nishida, 2007). In particular, the presence of several lineage-specific subclades as well as unresolved subclades generate novel hypotheses regarding TAAR subtype origins. While TAAR-2,3,4 form a well-resolved monophyly, sister to a clade containing subtypes TAAR-5,6,7,8,9, the latter clade is less straight-forward to interpret. Indeed, the TAAR subtypes -6, -7, and -9 do not constitute distinct monophyletic groups, suggesting either poor NCBI sequence annotation or rampant diversification within this subclade. If we assume that these NCBI annotations are reasonably correct, we can deduce that this clade's ancestral sequence was most similar to TAAR-7 and subsequently diversified independently into TAAR-9 and TAAR-6, which in turn gave rise to the monophyletic TAAR-8. Furthermore, lobe-finned (coelacanth) sequences do not clearly cluster with any TAAR subtypes, likely due to this lineage's ancient divergence and unique evolutionary trajectory of this lineage (Amemiya et al., 2013). The phylogenetic distribution of lobe-finned fish sequences may aid future endeavors to tease apart evolutionary origins of certain TAAR subtypes, specifically whether they represent teleost-specific duplications (Gloriam et al., 2005) or whether they represent ancient TAARs that emerged before teleost divergence but were secondarily lost in lobe-finned fish and/or tetrapods.

In addition, a small clade sister to TAAR (labeled in Figure 3 and Figure 4as TAAR$*$) strictly contains sequences annotated by NCBI as "5HTR4-like," which might suggest that 5HTR-4 is in fact paraphyletic, diversifying gradually before giving rise to TAARs. However, all sequences in TAAR$*$ belong taxonomically either to teleost or *Xenopus tropicalis*. Thus, we suspect that these sequences are actually misannotated, and therefore this clade in fact corresponds to the so-called TAAR V cluster identified by Hashiguchi and Nishida (2007). Indeed, Hashiguchi and Nishida (2007) showed that the TAAR V cluster is an outgroup to all other vertebrate TAAR sequences, and moreover that 5HTR-4 is an outgroup to the overall TAAR clade, as our phylogeny similarly shows.

Finally, we emphasize the limits of phylogenetic inference for understanding the complex evolutionary histories of expanding gene families. The majority of the deeper splits in our phylogeny have fairly weak bootstrap support, and thus this phylogeny alone is not sufficient for fully resolving the relationships among biogenic amine receptor classes. Indeed, modern phylogenetic inference methods are ill-suited for deducing such relationships, particularly because MSA gaps are treated simply as missing data. In reality, gaps represent the evolutionary events of insertion and deletions (indels). Unfortunately, current phylogenetic methods focus solely on the substitution process, effectively ignoring that gaps are indeed evolutionary events containing important information and ultimately hindering phylogenetic accuracy Morrison (2008); Loytynoja and Goldman (2008); Warnow (2012); Luan et al. (2013). Like most protein families, GPCRs do not simply diversify through nucleotide substitutions, as novel GPCRs tend to arise after major indel events in the ICL and ECL domains (Bockaert and Pin, 1999). Unfortunately, the evolutionary intermediates connecting sequences have long-since disappeared from genomes, and there is no obvious way to infer the sequences of these "missing links." Thus, additional approaches, notably syntenic analyses (Sundstrom et al., 2010; Widmark et al., 2011; Yegorov and Good, 2012; Hwang et al., 2013), combined with the phylogeny presented here should prove useful towards resolving the evolutionary history of vertebrate biogenic amine receptors.

CONCLUSIONS

In this paper, we have established a comprehensive, high-quality, structurally-curated MSA of vertebrate biogenic amine receptors. We hope that this MSA, along with its ML phylogeny, (freely available from `https://github.com/sjspielman/amine_r`
will serve as a robust resource for future studies investigating the evolutionary dynamics as well as structural/functional constraints operating within distinct receptor clades or indeed universal patterns that generally govern biogenic amine receptor evolution. Future work may seek to combine the analyses we have accomplished here with syntenic or molecular clock approaches to elucidate receptor origins and precise evolutionary trajectories. Moreover, our curated MSA should prove useful in increasing accuracy in homology modeling and/or pharmaceutical development for these clincally important receptors (Kristiansen, 2004; Ishiguro, 2004; Evers et al., 2005; Mason et al., 2012).

METHODS

Sequence Collection and Processing

We collected protein sequences using PSI-BLAST (Altschul et al., 1997), specifically from the RefSeq (v2.2.29+) database (Pruitt et al., 2013), for 42 distinct human biogenic amine receptor sequences representing the full range of known receptors in the human genome. To obtain distant yet well-supported orthologs, we ran each PSI-BLAST search for 5 iterations with a e-value cutoff of 10^{-20}, a sequence identity threshold of 25%, and a length difference of 50%. After combining all sequences recovered from the individual PSI-BLAST searches, we discarded duplicate sequences, leaving a total of 4232 PSI-BLAST results. We then filtered this sequence set by removing sequences from non-vertebrate taxa, sequences

annotated as low-quality, pseudogene, and/or partial, and sequences which contained more than 1% ambiguous (i.e. B, X, or Z) residues. We additionally removed any sequences that could not be robustly considered GPCRs. We used the program GPCRHMM (Wistrand et al., 2006) to determine whether a given sequence was indeed a GPCR, and we discarded sequences which had either a local or global GPCRHMM score less than 10, both extremely conservative thresholds. Thus, while it is possible that some true GPCRs were discarded, these high thresholds for both local and global scores provide a very high confidence that all retained sequences were indeed GPCRs. Together, these filters left a total of 3464 receptor sequences.

Sequence Alignment and Phylogenetic Reconstruction

Before aligning sequences, we used the program GPCRHMM (Wistrand et al., 2006) to assign each residue in all protein sequences to its respective structural domain (extracellular, transmembrane, or intracellular) using a 0.5 posterior probability cutoff. We then aligned and filtered sequences according to the strategy outlined in Figure 1, which specifically employed MAFFT v7.149b Katoh and Standley (2013).

All phylogenies were created using RAxML v8.1.1 Stamatakis (2014) using the LG+F (Le and Gascuel, 2008) amino acid exchangability matrix and the CAT model of site heterogeneity (Stamatakis, 2006), with the default 25 rate categories. For inferences incoroporating structural partitions, each partition was assigned a unique evolutionary model using these settings. Final parameter values for all phylogenetic inferences were optimized with the GAMMA model of heterogeneity. We performed 200 bootstrap replicates for each phylogeny.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant R01 GM088344, ARO grant W911NF-12-1-0390, DTRA grant HDTRA1-12-C-0007, and NSF Cooperative Agreement No. DBI-0939454 (BEACON Center). Computational resources were provided by the University of Texas at Austin's Center for Computational Biology and Bioinformatics (CCBB).

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:6:716–723.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402.

Amemiya, C. T., Alfoldi, J., Lee, A. P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J. J., Robinson, M., Dorrington, R. A., Gerdol, M., Aken, B., Biscotti, M. A., Barucca, M., Baurain, D., Berlin, A. M., Blatch, G. L., Buonocore, F., Burmester, T., Campbell, M. S., Canapa, A., Cannon, J. P., Christoffels, A., De Moro, G., Edkins, A. L., Fan, L., Fausto, A. M., Feiner, N., Forconi, M., Gamieldien, J., Gnerre, S., Gnirke, A., Goldstone, J. V., Haerty, W., Hahn, M. E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S. I., Kuraku, S., Lara, M., Levin, J. Z., Litman, G. W., Mauceli, E., Miyake, T., Mueller, M. G., Nelson, D. R., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C. P., Prohaska, S. J., Przybylski, D., Saha, N. R., Ravi, V., Ribeiro, F. J., Sauka-Spengler, T., Scapigliati, G., Searle, S. M. J., Sharpe, T., Simakov, O., Stadler, P. F., Stegeman, J. J., Sumiyama, K., Tabbaa, D., Tafer, H., Turner-Maier, J., van Heusden, P., White, S., Williams, L., Yandell, M., Brinkmann, H., Volff, J.-N., Tabin, C. J., Shubin, N., Schartl, M., Jaffe, D. B., Postlethwait, J. H., Venkatesh, B., Di Palma, F., Lander, E. S., Meyer, A., and Lindblad-Toh, K. (2013). The african coelacanth genome provides insights into tetrapod evolution. *Nature*, 496:311 – 316.

Anbazhagan, P., Purushottam, M., Kiran Kumar, H. B., Mukherjee, O., Jain, S., and Sowdhamini, R. (2010). Phylogenetic analysis and selection pressures of 5-HT receptors in human and non-human primates: Receptor of an ancient neurotransmitter. *Journal of Biomolecular Structure and Dynamics*, 27(5):581–598.

Bockaert, J. and Pin, J. P. (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *The EMBO Journal*, 18(7):1723–1729.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol Method Res*, 33:261–304.

Callier, S., Snapyan, M., Crom, S., Prou, D., Vincent, J.-D., and Vernier, P. (2003). Evolution and cell biology of dopamine receptors in vertebrates. *Biology of the Cell*, 95(7):489–502.

Evers, A., Hessler, G., Matter, H., and Klabunde, T. (2005). Virtual screening of biogenic amine-binding G-protein coupled receptors: Comparative evaluation of protein- and ligand-based virtual screening protocols. *Journal of Medicinal Chemistry*, 48(17):5448–5465.

Franzosa, E., Xue, R., and Y, X. (2013). Quantitative residue-level structure-evolution relationships in the yeast membrane proteome. *Genome Biol Evol*, 5:734–744.

Fredriksson, R., Lagerstrom, M., Lundin, L., and Schioth, H. (2003). The G-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology*, 63(6):1256–1272.

Gloriam, D. E. I., Bjarnadóttir, T. K., Yan, Y.-L., Postlethwait, J. H., Schioth, H. B., and Fredriksson, R. (2005). The repertoire of trace amine G-protein-coupled receptors: large expansion in zebrafish. *Molecular Phylogenetics and Evolution*, 35(2):470–482.

Hashiguchi, Y. and Nishida, M. (2007). Evolution of trace amine associated receptor (taar) gene family in vertebrates: Lineage-specific expansions and degradations of a second class of vertebrate chemosensory receptors expressed in the olfactory epithelium. *Mol Biol Evol*, 24(9):2099–2107.

Hwang, J. I., Moon, M. J., Park, S., Kim, D. K., Cho, E. B., Ha, N., Son, G. H., Kim, K., Vaudry, H., and Seong, J. Y. (2013). Expansion of secretin-like G protein-coupled receptors and their peptide ligands via local duplications before and after two rounds of whole-genome duplication. *Mol Biol Evol*, 30(5):1119–1130.

Ishiguro, M. (2004). Ligand-binding modes in cationic biogenic amine receptors. *ChemBioChem*, 5(9):1210–1219.

Jordan, G. and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*, 29:1125–1139.

Julenius, K. and Pedersen, A. G. (2006). Protein evolution is faster outside the cell. *Mol Biol Evol*, 23:2039–2048.

Kakarala, K. K. and Jamil, K. (2014). Sequence-structure based phylogeny of GPCR class A rhodopsin receptors. *Molecular Phylogenetics and Evolution*, 74(C):66–96.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*, 30:772–780.

Krishnan, A., Almén, M. S., Fredriksson, R., and Schioth, H. B. (2012). The origin of gpcrs: Identification of mammalian like rhodopsin, adhesion, glutamate and frizzled gpcrs in fungi. *PLoS ONE*, 7(1):e29817.

Kristiansen, K. (2004). Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacology & Therapeutics*, 103(1):21–80.

Lagerstrom, M. C. and Schioth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov*, 7:339–357.

Le, S. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol*, 25:1307–1320.

Lindemann, L., Ebeling, M., Kratochwil, N. A., Bunzow, J. R., Grandy, D. K., and Hoener, M. C. (2005). Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. *Genomics*, 85(3):372–385.

Loytynoja, A. and Goldman, N. (2008). Phylogeny–aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320:1632–1635.

Luan, P.-t., Ryder, O. A., Davis, H., and Zhang, Y.-p. (2013). Incorporating indels as phylogentic characters: impact for interfamilial relationships within Arctoidea (Mammalia: Carnivora). *Molecular Phylogenetics and Evolution*, 66:748 – 756.

Mason, J. S., Bortolato, A., Congreve, M., and Marshall, F. H. (2012). New insights from structural biology into the druggability of G protein-coupled receptors. *Trends in Pharmacological Sciences*, 33(5):249–260.

Morrison, D. A. (2008). A framework for phylogenetic sequence alignment. *Plant Systematics and Evolution*, 282(3-4):127–149.

Oberai, A., Joh, N. H., Pettit, F. K., and Bowie, J. U. (2009). Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proc Natl Acad Sci USA*, 106:17747–17750.

Ogden, T. H. and Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, 55(2):314–328.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, 340:385–395.

Pei, J., Kim, B. H., and Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nuc Acids Res*, 36(7):2295–2300.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., Tully, R. E., Wallin, C., Webb, D., Weber, J., Wu, W., DiCuccio, M., Kitts, P., Maglott, D. R., Murphy, T. D., and Ostell, J. M. (2013). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*, 42(D1):D756–D763.

Rompler, H., Staubert, C., Thor, D., Schulz, A., Hofreiter, M., and Schöneberg, T. (2007). G protein-coupled time travel. evolutionary aspects of GPCR research. *Molecular Interventions*, 7(1):17–25.

Schoneberg, T., Schulz, A., Biebermann, H., Hermsdorf, T., H, R., and Sangkuhl, K. (2004). Mutant G protein-coupled receptors as a cause of human diseases. *Pharmacol Ther*, 104:173–206.

Spielman, S. J. and Wilke, C. O. (2013). Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. *J Mol Evol*, 76(3):172–182.

Stamatakis, A. (2006). Phylogenetic models of rate heterogeneity: a high performance computing perspective. In *Proc. of IPDPS2006*.

Stamatakis, A. (2014). RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312 – 1313.

Staubert, C., Böselt, I., Bohnekamp, J., Rompler, H., Enard, W., and Schöneberg, T. (2010). Structural and functional evolution of the trace amine-associated receptors TAAR3, TAAR4 and TAAR5 in primates. *PLoS ONE*, 5(6):e11133.

Stäubert, C., Le Duc, D., and Schöneberg, T. (2013). Examining the dynamic evolution of G protein-coupled receptors. pages 23–43. Humana Press, Totowa, NJ.

Stevens, T. J. and Arkin, I. T. (2001). Substitution rates in alpha-helical transmembrane proteins. *Prot Sci*, 10:2507–2517.

Sundstrom, G., Dreborg, S., and Larhammar, D. (2010). Concomitant duplications of opioid peptide and receptor genes before the origin of jawed vertebrates. *PLoS One*, 5:e10512.

Tourasse, N. J. and Li, W.-H. (2000). Selective constraints, amino acid composition, and the rate of protein evolution. *Mol Biol Evol*, 17:656–664.

Vassilatis, D. K. et al. (2003). The G protein-coupled receptor repertoires of human and mouse. *Proc Natl Acad Sci USA*, 100(8):4903–4908.

Warnow, T. (2012). Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Curr*, 4:RRN1308.

Widmark, J., Sundstrom, G., Ocampo, D., and Larhammar, D. (2011). Differential evolution of voltage-gated sodium channels in tetrapods and teleost fishes. *Mol Biol Evol*, 28:859—-871.

Wistrand, M., Käll, L., and Sonnhammer, E. L. L. (2006). A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Prot Sci*, 15(3):509–521.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476.

Yamamoto, K., Mirabeau, O., Bureau, C., Blin, M., Michon-Coudouel, S., Demarque, M., and Vernier, P. (2013). Evolution of dopamine receptor genes of the D1 class in vertebrates. *Mol Biol Evol*, 30(4):833–843.

Yegorov, S. and Good, S. (2012). Using paleogenomics to study the evolution of gene families: Origin and duplication history of the relaxin family hormones and their receptors. *PLoS ONE*, 7(3):e32923.

FIGURES AND TABLES

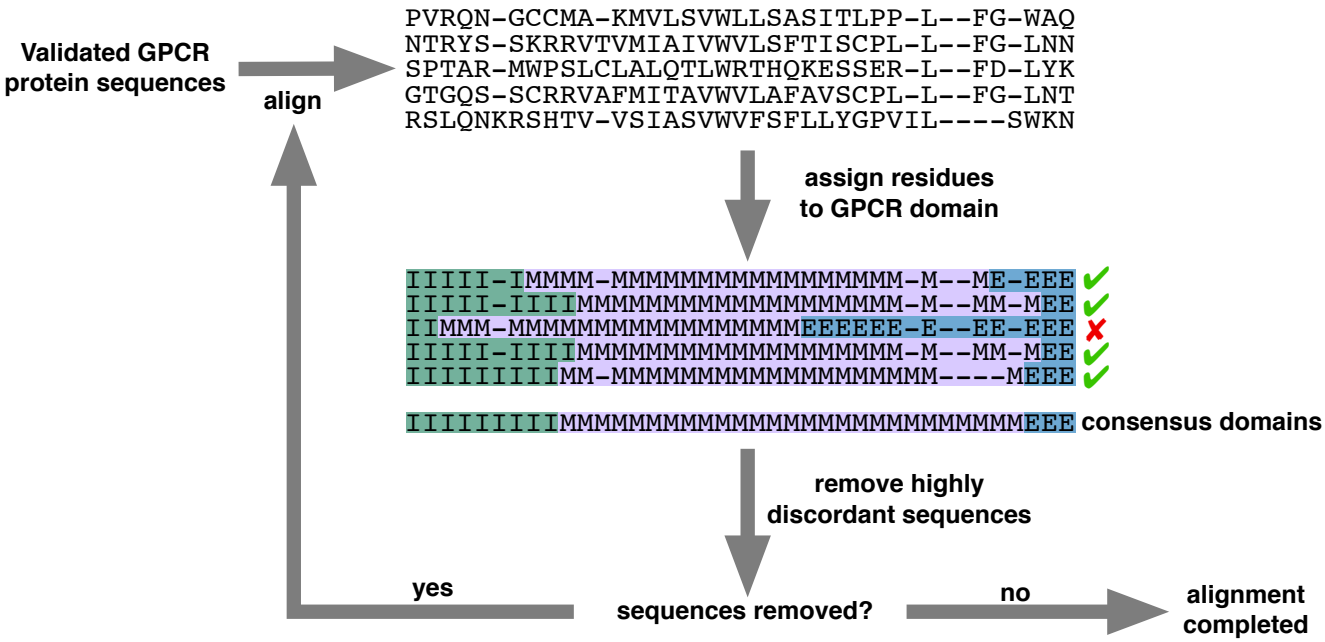

Figure 1. Iterative alignment strategy to create a structurally-curated MSA of vertebrate biogenic amine receptors. A total of 3464 sequences were initially input ("Validated GPCR protein sequences"), and the final MSA contained 3039 protein sequences. Residues marked with "I" represent intracellular residues, those marked with "M" represent transmembrane residues, and those marked with "E" represent extracellular residues. MSA gaps were regarded as missing data when determining each column's consensus structural domain. Sequenced were removed ("remove highly discordant sequences") if $\geq 5\%$ of columns belonged to a different structural domain than the respective consensus domain.

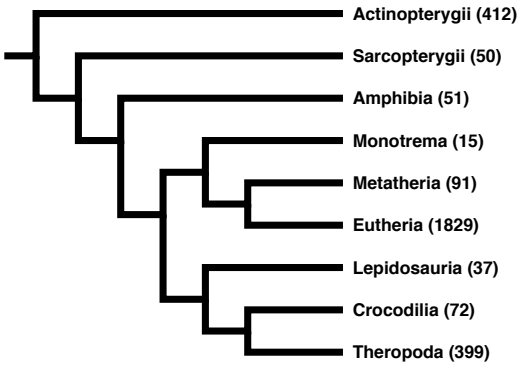

Figure 2. Cladogram of the taxonomic distribution of all sequences in the final structurally-curated MSA. All sequences belonged to the *Euteleostomi* clade of jawed vertebrates. Numbers in parentheses indicate the total number of biogenic amine receptors in the respective clade. We note that our MSA is particularly enriched for sequences from Eutherian (placental mammal) species, likely due to the stringent filters we applied to sequence collection which favored fully-sequenced genomes.

MSA	Partitions?	$\ln L$	k	ΔAIC
Structural Masked	Yes	-505500.8	6115	0
Structural Masked	No	-515991.7	6095	1752
Structural Unmasked	Yes	-515343.6	6115	19685
Structural Unmasked	No	-515991.7	6095	20941
Naive	No	-589703.7	6945	170047

Table 1. ΔAIC scores relative to the best performing for phylogenies. The column labeled "Partitions?" indicates whether phylogenetic inference was conducted with distinct TM (transmembrane) and EM (extramembrane) partitions. AIC is computed as $\text{AIC} = 2(k - \ln L)$, where k is the number of free parameters of the model, and $\ln L$ is the log-likelihood (Akaike, 1974; Burnham and Anderson, 2004). AIC scores are reported here relative to the phylogeny with the lowest AIC score (structural masked with partitions), representing the best-fitting phylogeny.

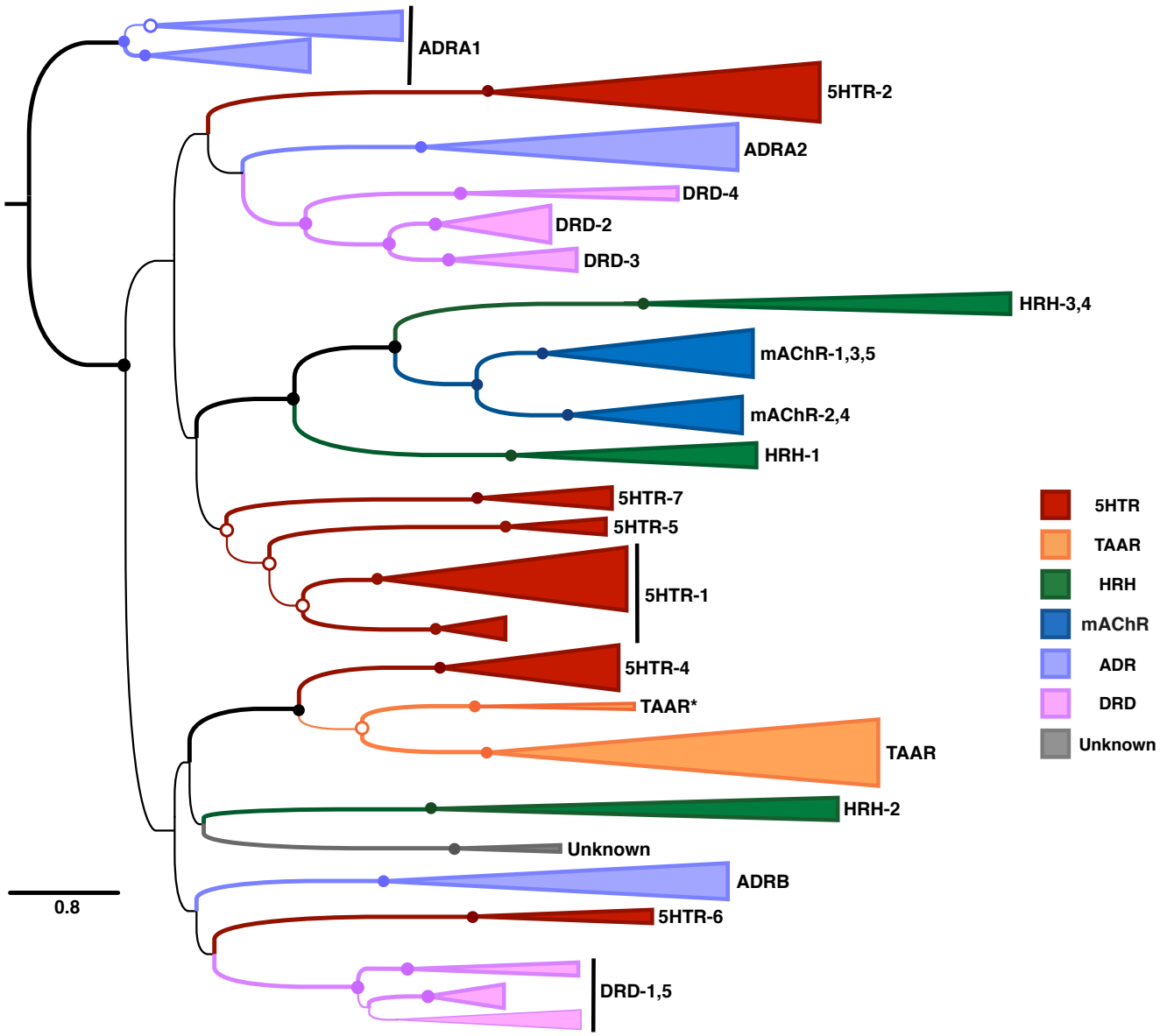

Figure 3. Maximum likelihood phylogeny of vertebrate biogenic amine receptors built using the masked structural MSA in RAxML. Nodes with open circles indicate $\geq$ 50% bootstrap support, and nodes with closed circles and thick lines indicate $\geq$ 90% bootstrap support. Bioaminergic receptors are abbreviated as 5HTR, serotonin receptors; TAAR, trace amine-associated receptors; HRH, histamine receptors; mAChr, muscarinic acetylcholine receptors; ADR, adreneric receptors; and DRD, dopamine receptors. The clade labeled "Unknown" could not be clearly identified as one of the major receptor types and may represent a novel biogenic amine receptor clade.

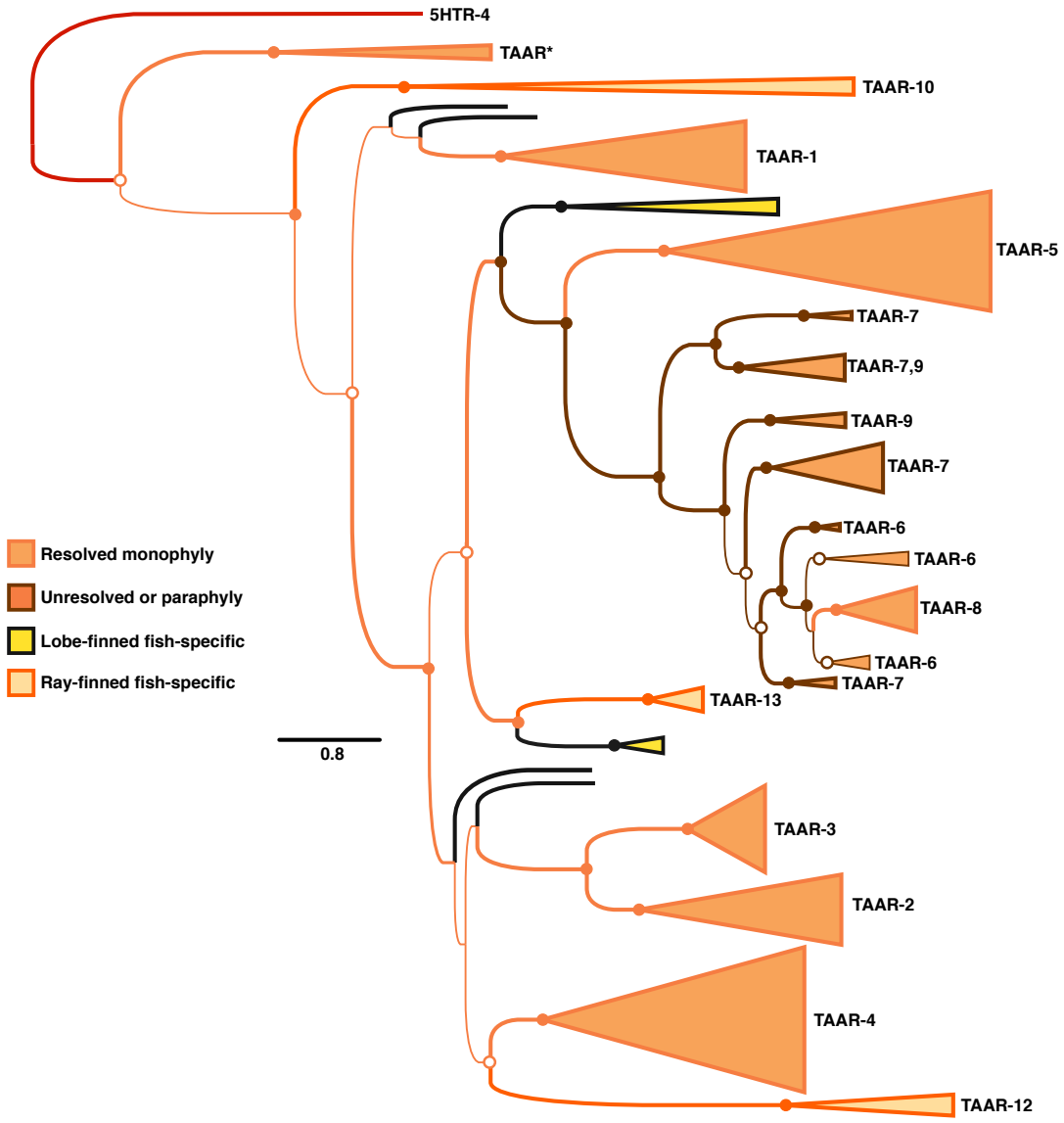

Figure 4. Subclade of the TAAR receptors within the phylogeny shown in Figure 3. Nodes with open circles indicate $\geq 50\%$ bootstrap support, and nodes with closed circles and thick lines indicate $\geq 90\%$ bootstrap support.