Ramakrishnan Varadarajan et al. BMC Evolutionary Biology (2018) 18:1
DOI 10.1186/s12862-017-1072-2

BMC Evolutionary Biology

DATABASE Open Access

Genome-wide protein phylogenies for four African cichlid species

Ajay Ramakrishnan Varadarajan, Rohini Mopuri, J. Todd Streelman and Patrick T. McGrath*

Abstract

Background: The thousands of species of closely related cichlid fishes in the great lakes of East Africa are a powerful model for understanding speciation and the genetic basis of trait variation. Recently, the genomes of five species of African cichlids representing five distinct lineages were sequenced and used to predict protein products at a genome-wide level. Here we characterize the evolutionary relationship of each cichlid protein to previously sequenced animal species.

Results: We used the Treefam database, a set of preexisting protein phylogenies built using 109 previously sequenced genomes, to identify Treefam families for each protein annotated from four cichlid species: *Metriaclima zebra*, *Astatotilapia burtoni*, *Pundamilia nyererei* and *Neolamporologus brichardi*. For each of these Treefam families, we built new protein phylogenies containing each of the cichlid protein hits. Using these new phylogenies we identified the evolutionary relationship of each cichlid protein to its nearest human and zebrafish protein. This data is available either through download or through a webserver we have implemented.

Conclusion: These phylogenies will be useful for any cichlid researchers trying to predict biological and protein function for a given cichlid gene, understanding the evolutionary history of a given cichlid gene, identifying recently duplicated cichlid genes, or performing genome-wide analysis in cichlids that relies on using databases generated from other species.

Background

The rapid decrease in sequencing costs and the development of broadly applicable genetic tools like TALENs and CRISPR/Cas9 has facilitated the development of a large number of new species as model organisms [10, 20, 26, 42]. For evolutionary biologists, this has been especially fruitful – species with unique evolutionary traits can now be used as model organisms to identify and understand the underlying genetic and cellular mechanisms responsible for trait changes [16]. For example, threespine sticklebacks have long fascinated evolutionary biologists for their coexisting phenotypically divergent forms including freshwater/anadromous pairs [18, 39, 40]. Freshwater lakes created after the retreat of Pleistocene glaciers have been populated by marine sticklebacks, evolving repeated changes in a number of traits. These adaptations include morphological changes to body shape, pigmentation changes, salt handling, and reproductive related behaviors [3, 39]. A combination of quantitative genetics and resequencing of individuals isolated from freshwater and saltwater habitats identified a large number of loci putatively responsible for evolution of marine-freshwater ecotypes [5, 7, 25]. An important conclusion from this research, and a number of other individual examples [38], is that despite the large number of genes that control a trait, natural selection can act in predictable ways, isolating genetic changes in preferred genes in response to specific environmental shifts. An important goal now is to identify additional examples of repeated evolution, and understand why particular genes are repeatedly selected.

Cichlid fishes offer an attractive avenue for this type of research. Cichlids are well-known for their adaptive radiations in the Great Lakes of East Africa. The three largest radiations in Lakes Victoria, Lake Malawi, and Lake Tanganyika have generated between 250 and 500 species per lake in a period of time that ranges from 100,000 to 12 million years [4, 30]. These radiations resulted in exceptional phenotypic diversity in behavior, neurodevelopment, body shape, sexual traits, and ecological

* Correspondence: patrick.mcgrath@biology.gatech.edu
Department of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Dr., Atlanta, GA 30332, USA

specialization. However, due to the speed of evolution, nucleotide diversity between these species is on the order of nucleotide diversity within the human population [4, 34]. Further, genetic barriers have not formed in this short period, allowing for genetics - phenotypically-divergent species can still interbreed. These peculiarities of the cichlid family make genomics and quantitative genetics approaches particularly attractive. Genes responsible for phenotypic diversity can be identified using quantitative mapping approaches in progeny of intercrossed species, association mapping in outbred animals, or tissue-specific transcriptomics in behaving animals. To facilitate these approaches, high-quality genomes for five cichlid fishes were generated [4]. It is anticipated that genetic variants and genes responsible for a variety of interesting trait differences will be identified in the coming years.

Due to the difficulty of experimental study of cichlids in the laboratory, assignment of molecular and biological function to genes relies almost exclusively on homology to proteins characterized biochemically, or in model organisms such as *Caenorhabditis elegans, Drosophila melanogaster, Danio rerio,* or *Mus musculus*. Homologous proteins share a common evolutionary ancestry [14], suggesting shared biochemical and/or biological role, justifying the use of homology to assign function to genes identified in cichlid fish. Proteins with shared homology can be characterized as orthologs (which diverged from a common ancestor due to speciation) or paralogs (which diverged from a common ancestor due to a gene duplication event). In general, orthologs are expected to retain similar (if not identical) function with each other. Paralogs are expected to acquire novel function and/or biological roles [31]. This hypothesis has been called the ortholog conjecture [41]. While conceptually attractive, the empirical support for this conjecture is still under debate. [2, 6, 15, 32, 41]. For cichlids, paralogs are thought to be especially relevant to their evolution - the cichlid lineage has undergone an increased rate of gene duplication, suggesting that these novel genes could serve important roles in the cichlid's adaptive radiations [4, 35]. Cichlids also belong to the teleost infraclass of fish, whose ancestors have undergone a genome-wide duplication event resulting in the duplication of a large number of genes [48]. Gene duplication can allow resolution of adaptive conflict by allowing a bifunctional ancestral gene to resolve into two specialized genes [36]. These gene duplicates have been proposed to play a role in the evolutionary success of the teleost fish, which make up ~96% of all fish. Phylogenetic relationships could potentially be used to identify the cichlid genes that have undergone subfunctionalization. For all of these reasons, it would be helpful to place each cichlid protein into a phylogeny to aid in predicting the gene function for a given cichlid gene.

In this report, we utilized the TreeFam database of protein phylogenies to create protein phylogenies for all completely sequenced cichlid genomes. We analyzed these phylogenies to determine evolutionary relationships for each of these cichlid genes. This data is available for download or searching on a web server (http://cichlids.biosci.gatech.edu/), and should be useful to any researchers studying cichlid fish.

Construction and content

Overview

We employed a phylogeny-based approach to study the function and evolution of 97,862 proteins taken from four East African cichlid species. Nile tilapia (*O. niloticus*) is already included in the TreeFam database. Our aim was to assign each cichlid gene to a pre-defined gene family to identify homologous proteins and their evolutionary relationship. To accomplish this, we used TreeFam, a database of phylogenetic trees drawn from 109 animal genomes [33, 43, 45]. A webserver implementing the TreeFam pipeline is provided (www.treefam.org) to add new proteins of interest to existing TreeFam trees. We implemented this pipeline locally to perform this on a genomic basis.

Datasets and TreeFam analysis

Protein coding sequences and annotation files for four cichlid species, *A. burtoni*, *M. zebra*, *N. brichardi*, and *P. nyererei,* were obtained from the supplemental dataset from the genome sequencing paper [4]. An improved genome for *M. zebra* was also recently published; protein coding sequence and genome annotation files from this paper were downloaded from NCBI [8]. Annotation files were parsed using custom Python scripts and used to identify the longest protein isoform and amino acid sequence for each gene. This was done to limit the phylogeny to one representative protein isoform for each gene. To assign each of these proteins to a single TreeFam family, we utilized the treefamscan.pl script provided as part of the TreeFam API [45]. This script uses the program HMMER to identify matches using hidden Markov model profiles generated for each of the TreeFam families [11]. After this had run on all of the proteins, we collected all of the protein sequences that best matched a given TreeFam to add these to the preexisting phylogeny. Multiple sequence alignments and phylogenies for each TreeFam were retrieved from a locally cloned SQL database with API utilities provided by TreeFam. We used MAFFT (version 7.221) to add the new cichlid proteins to the retrieved multiple sequence alignment using the –add, –reorder, and -anysymbol options [28, 29]. The aligned output was then used to add the new proteins to the retrieved phylogeny file using RaXML (version 8.1.15) using the GAMMA model for rate heterogeneity with the WAG substitution matrix [46, 47, 49]. These options were chosen to remain consistent with the approach taken by TreeFam [45]. Support values for the new nodes were taken from the maximum likelihood files created

by RaXML and added to the phylogenies using Genesis (https://github.com/lczech/genesis).

Identification of closest relationships to human and zebrafish proteins

For each cichlid protein, we used custom Python scripts to identify the closest human and zebrafish protein using the phylogenetic tree produced by RAxML. The structures of each tree were analyzed using the ETE toolkit, which provides a Python framework for analysis and visualization of protein trees [22]. Trees were rooted using a midpoint outgroup method implemented by the get_midpoint_outgroup function. To find the closest human protein and its evolutionary relationship with a cichlid protein of interest, the trees were then traversed to identify the smallest subtree containing the cichlid protein and one or more human protein. If such a subtree could not be found (i.e. there was no human protein in the phylogeny), the relationship was defined as NoHomolog. If the subtree contained one or more human proteins, the ortholog/paralog relationship was determined using the species overlap algorithm [21] implemented in the ETE toolkit. If the subtree contained multiple human proteins, the closest human protein to the cichlid protein of interest was identified using the shortest branch length. If the relationship between the human protein and the cichlid protein was orthologous, we also determined the one-to-X relationship and the median number of hits in all cichlid genomes. This could be useful for researchers interested in studying teleost-specific duplications or duplications that occurred in the cichlid lineage. To convert the Ensembl protein ID's of the human proteins to HGNC identifiers [17], we downloaded mapping data from Ensembl BioMart [1]. An essentially identical process was also performed between all cichlid proteins with zebrafish proteins. An excel spreadsheet (one per species) was then created for each cichlid gene for this information.

PDFs of the resulting phylogenies were rendered using the ETE toolkit. A full size version of each TreeFam phylogeny was created using all species. In addition, a smaller PDF was created from a pruned tree containing a limited number of well-characterized species (human (*H. sapiens*), mouse (*M. musculus*), zebrafish (*D. rerio*), fruit fly (*D. melanogaster*), and nematode (*C. elegans*)), the closely related Nile tilapia (*O. niloticus*), and the four new cichlid species.

Utility and discussion

Identification of human and zebrafish relationships for each cichlid gene

The cichlids species of East Africa have become a popular genomic model to understand the evolution of a number of traits, including differences in morphology, coloration and behavior. To broaden our understanding of the function and evolutionary history of the genes that are encoded in the genomes of four recently-sequenced cichlid species, we performed phylogenetic analysis using the previously published TreeFam pipeline to add the new cichlid proteins to preexisting protein phylogenies generated from a large number of animal species (Fig. 1). The most current version of the TreeFam database [45], which contains 15,736 phylogenetic trees generated from 109 animal genomes covering ~2.2 million sequences, can be used to study evolutionary relationships between homologous proteins. While this database already includes the African cichlid *O. niloticus* (Nile tilapia), it does not contain four recently sequenced African cichlids: *M. zebra* from Lake Malawi, *P. nyererei* from Lake Victoria, *N. brichardi* from Lake Tanganyika, and *A. burtoni* found in a variety of African lakes and rivers. For all four cichlid species, the majority of cichlid genes, 82.2% – 84.7%, contained a hit to a preexisting TreeFam family (Fig. 2). Using the resulting phylogenies, we

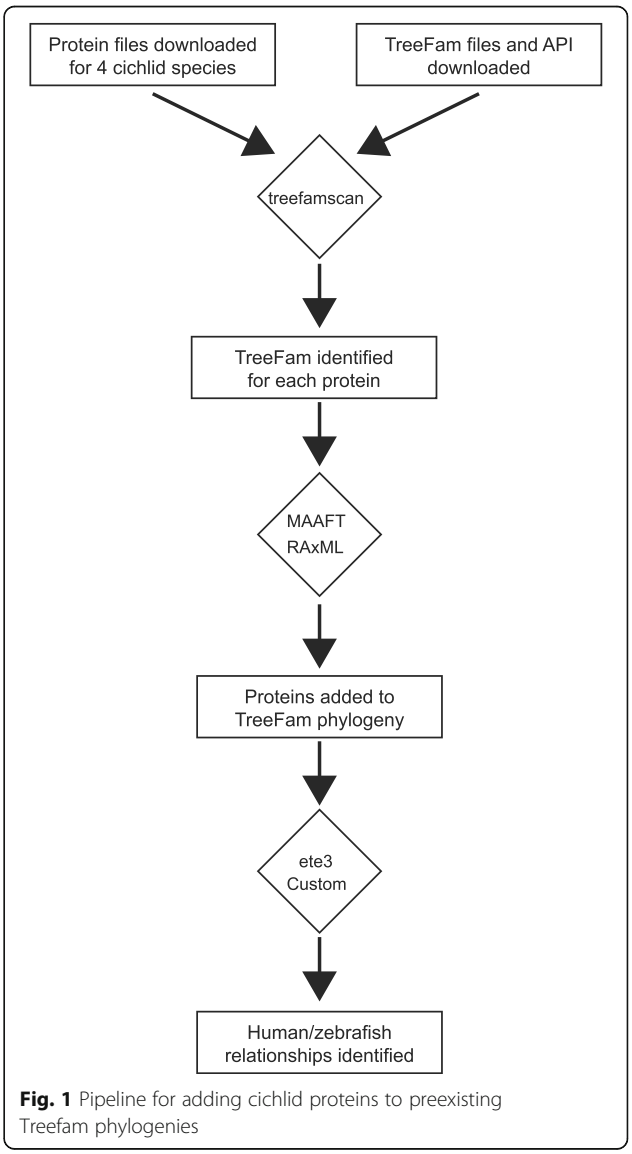

Fig. 1 Pipeline for adding cichlid proteins to preexisting Treefam phylogenies

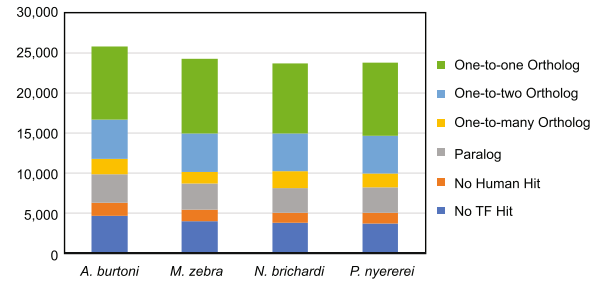

Fig. 2 Summary of the human evolutionary relationships found in each species. One-to-two ortholog indicates a gene potentially duplicated and retained in an ancient teleost ancestor

identified the closest human and zebrafish gene along with the evolutionary relationship to the cichlid. These included traditional evolutionary relationships (Ortholog and Paralog) and one-to-many relationships to account for the large number of cichlid genes that duplicated either in the ancestral teleost lineage and are retained in the extant species.

Data accessibility

This data is intended as a resource for the cichlid community. We have provided access to this data in three ways. 1. Two PDF files for each TreeFam were generated for the purposes of human inspection. One PDF contains a phylogeny for a TreeFam from a limited number of species: humans, four well-characterized model organisms (*C. elegans, D. melanogaster, D. rerio,* and *M. musculus*), Nile tilapia (*O. niloticus*), and the four recently studied cichlid species. The second PDF contains a phylogeny of all species used in the analysis. While the second phylogeny is the most complete, it is difficult to analyze due to the large number of species. This data is hosted on a web server (http://cichlids.biosci.gatech.edu/)and can be searched using cichlid gene names, TreeFam IDs, or human and zebrafish names. 2. Excel files for each cichlid species that contain each gene, its best hit to a human and zebra fish gene, and its evolutionary relationship to that gene. We anticipate this data will be useful for genomic scale analysis. For example, the excel file can be loaded into scripts to automatically map cichlid genes to human or zebrafish homologs. This could be useful for the purposes of pathway analysis (such as gene ontology), which often are limited to human genes. 3. Finally, alignments and phylogenies of each TreeFam are available for download in Newick tree format. These will be useful for any researchers interested in automated analysis of the phylogenies for the purpose of enhancing the evolutionary relationships that we have reported here. For example, researchers could use this dataset to identify genes whose protein phylogenies contradict the species phylogenies.

Example phylogeny generated from a tree containing members of the TGFβ superfamily

To illustrate these evolutionary relationships as well as common issues users should be aware of in using these trees, we have included two figures of new phylogenies generated in this analysis. Figure 3 shows a subtree of TF351789, which includes members of the TGFβ-superfamily of proteins including BMP2 and BMP4. These proteins are ubiquitous throughout metazoans, and control proliferation and differentiation of cells throughout development [44]. This tree includes both ortholog and paralog relationships. For example, the subtree indicated by **a** in Fig. 3 shows one-to-one ortholog relationships between the cichlid proteins and human BMP4. These genes likely play similar biological roles in cichlids. Similarly, subtree **b** contains cichlid one-to-one orthologs to human BMP2 (with the exception of *M. zebra*, which will be discussed below), suggesting these genes play similar biological roles as the orthologs play in other species. There is also a cichlid-specific set of paralogs to BMP2 and BMP4 not present in *D. rerio* (subtree **c**) suggesting that there was a duplication of BMP2 or BMP4 in a recent common ancestor of all cichlid species following separation from the zebrafish lineage. It is not obvious from the phylogeny what biological role these genes might play. This clade of genes is potentially of interest to cichlid biologists, as they could play a role in the extensive morphological diversity observed among cichlid species. However, analysis of the full tree indicated that this clade contains genes from a large number of additional teleost fish along with a coelacanth fish (*L. chalumnae*) and an anole lizard (*A. carolinensis*) (Additional file 1: Figure S1). Further, blasting the protein sequence encoded by the ab.gene.s112.4 from *A. burtoni* to the *D. rerio* genome identified a match to a known protein annotated as BMP16 [13]. BMP16 does not appear to be present in the Treefam database, which explains why it was not present in the phylogeny. This set of BMP2/BMP4 paralogs thus seems to be a duplication that occurred in an ancient vertebrate ancestor of these fish (preceding the teleost ancestor) and lost in most tetrapod lineages as proposed by Marques et al. [37].

We observed a similar issue in the phylogeny surrounding the human IRX1 gene (TF319371) (Additional file 2: Figure S2). The Treefam phylogeny suggests that *D. rerio* contained a single ortholog to this gene while each of the cichlid species contained two copies of this gene. However, previous publications demonstrate that there are also two versions of IRX1 in *D. rerio* (called *irx1a* and *irx1b*) [9, 12]. Inspection of the Treefam data indicates that *irx1a* isn't present in the starting dataset. These examples illustrate a common issue to most genomic analysis. Since Treefam relies on genomic-scale predictions, there are likely errors within the resulting phylogenies. Users would do well to manually verify or repeat any of these phylogenies for genes they are especially interested in.

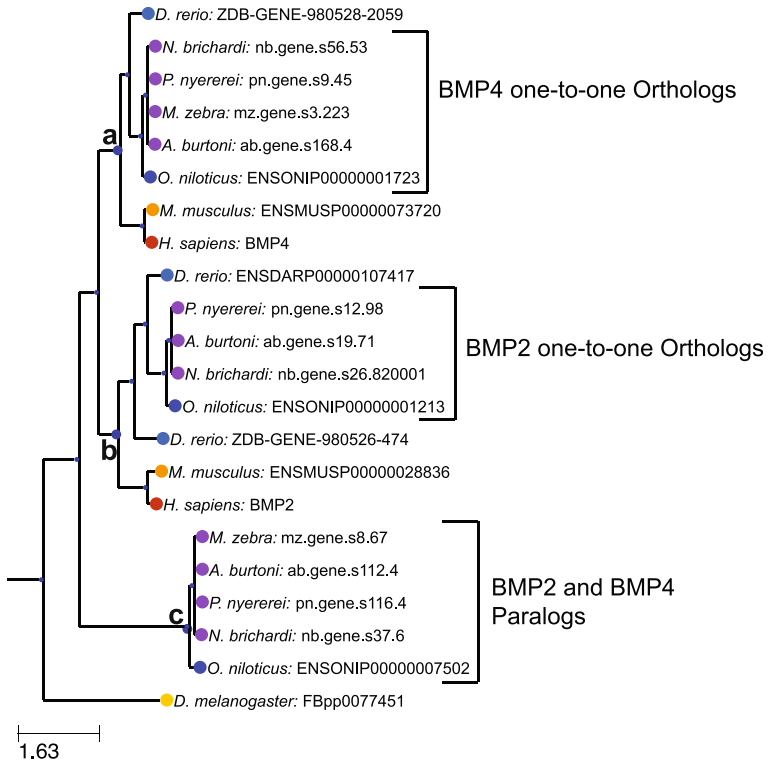

Fig. 3 Subtree from the TF351789 family from a limited number of cichlid species and well-studied model organisms. This family contains a number of BMP growth factors belonging to the transforming growth factor beta family. Letters indicate additional subtrees discussed in the text

We also were curious about the lack of a clear ortholog to BMP2 in *M. zebra* (Fig. 3). It seemed unlikely that this species could lose this protein entirely due to its essential function in bone development. We were able to track down this discrepancy to an error in the annotation file for *M. zebra*. Through blastp, we were able to identify mz.gene.s5.238 as a gene containing a strong match to BMP2. mz.gene.s5.238, however, was assigned to the TF314677 family, and predicted to be an ortholog to the human protein FERMT1. When we invested the protein sequence more closely, it became clear that mz.gene.s5.238 appeared to contain a fusion of two genes: an ortholog to BMP2 and an ortholog to FERMT1. Due to the longer length of FERMT1, mz.gene.s5.238 was assigned to the TreeFam containing FERMT1. This is unlikely to represent a real gene fusion, and the improved version of the *M. zebra* genome predicts separate gene products consistent with other species[8]. We observed a similar potential error with the PTGFR prostaglandin receptor. An ortholog of PTGFR has recently been shown to control female reproductive behaviors in the cichlid *A. burtoni* [27], however, the Treefam containing the human PTGFR gene (TF324982), did not contain an ortholog of this gene in *A. burtoni*. Again, this seems to be due to an error in annotation incorrectly predicting a fusion between two genes. The best blastp match ab.gene.s495.12 contains a fusion between two genes, an ortholog to PTGFR and an ortholog to the ZFYVE9. Due to the longer length of the ZFYVE9 protein, the ab.gene.s495.12 gene is assigned to the Treefam containing the human ZFYVE9. Again, this is unlikely to represent a real fusion, and it since has been corrected in new annotations. These two examples illustrate how errors in the gene annotation can lead to incorrect phylogenies.

Example phylogeny generated from a tree containing arginine vasopressin receptors

Figure 4 shows a subtree of the phylogeny for TF106499, which contains a number of receptors for the arginine vasopressin and oxytocin neuropeptides that are thought to play a role in social behavior and sexual motivation [19, 23]. We have limited this phylogeny to the clade containing the AVPR1A and AVPR1B human proteins. The clade indicated by **a** demonstrates a one-to-two ortholog relationship (Fig. 4). All of the sequenced cichlid species (along with zebrafish and other teleost fish) contain two genes that fall within this clade. This phylogeny suggests that the function of the ancestral AVPR1A gene bifurcated into two genes in an ancestor to the teleost lineage. While the phylogeny suggests that both of these receptors should retain a molecular role in arginine vasopressin/oxytocin signaling, the biological function of AVPR1A should not be assigned to either of the two genes in each cichlid species. Rather, experiment will be necessary to

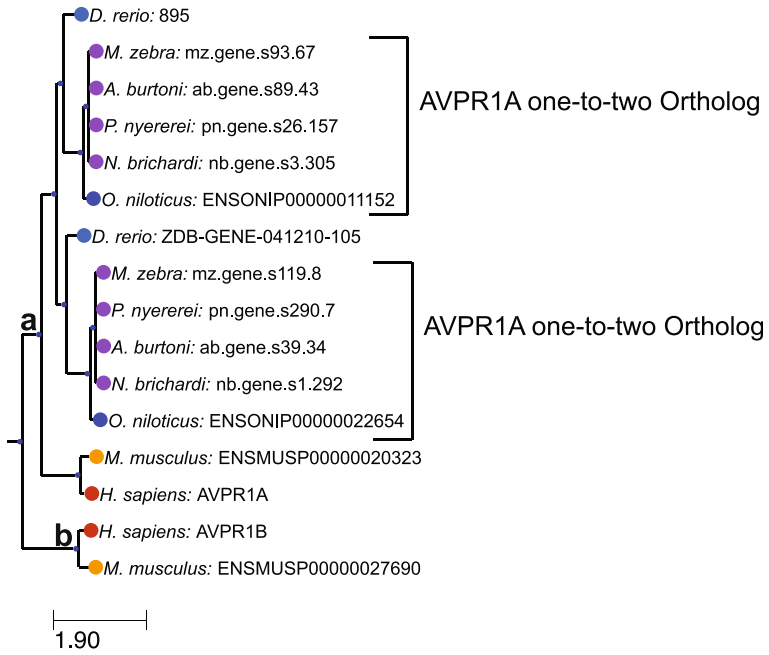

Fig. 4 Subtree from the TF106499 family from a limited number of cichlid species and well-studied model organisms. This family contains a number of G-protein receptors for the arginine vasopressin and oxytocin nonapeptide hormones. Letters indicate additional subtrees discussed in the text

parse out the biological function of each of these two orthologs. A recent paper characterizing the expression pattern of these two receptors in zebrafish demonstrated that these two genes are expressed in similar but non-overlapping cell types [24]. This phylogeny also contains the human AVPR1B protein. While mouse contains a clear ortholog to this gene, none of the cichlid species nor zebrafish contain an ortholog to this gene. Analysis of the full phylogeny suggests that AVPR1B was lost in the teleost fish completely. Thus, the phylogeny indicates that the biological functions assigned to AVPR1B through the study of mouse and other mammals should not be directly assigned to any of the cichlid homologs without experimental study.

Conclusions

This study reports a set of protein phylogenies generated for four recently sequenced African cichlids. We hope that these phylogenies will be useful for cichlid researchers for the purpose of inferring biological and molecular function of cichlid genes.

Additional files

Additional file 1: Figure S1. Full tree for the TF351789 family from all 109 species included in the Treefam database. This family contains a number of BMP growth factors belonging to the transforming growth factor beta family. (PDF 3001 kb)

Additional file 2: Figure S2. Subtree from the TF319371 family from a limited number of cichlid species and well-studied model organisms. This family contains a number of Iroquois-family of homeodomain transcription factors involved in patterning and other development processes. (PDF 1170 kb)

Acknowledgements

We thank Lucas Czech for use of his Genesis toolkit, Mateus Patricio for information on Treefam, and members of the McGrath and Streelman labs for comments on this work. This work was supported by NIH grants R21AG050304, R01GM114170, and the Ellison Medical Foundation (to P.T.M.) and NIH grants 1R01GM101095 and 2R01DE019637 (to J.T.S.).

Availability of data and materials

Supporting tables are included as supporting files. All data is available for searching or downloading at http://cichlids.biosci.gatech.edu.

Authors' contributions

ARV participated in data analysis and designed the web server. RM participated in data analysis and designing the web server. JTS participated in the design of the study. PTM conceived of the study, constructed the database, and drafted the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests

Publisher's Note

Received: 10 May 2017 Accepted: 15 November 2017
Published online: 08 January 2018

References

1. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, Garcia Giron C, Hourlier T, Howe K, Kahari A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel JH, White S, Zadissa A, Flicek P, Searle SM. The Ensembl gene annotation system. Database (Oxford). 2016;2016

2. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol. 2012;8(5):e1002514.
3. Bell MA, Foster SA. The evolutionary biology of the threespine stickleback. In: Oxford. New York: Oxford University Press; 1994.
4. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E, Turner-Maier J, Johnson J, Alcazar R, Noh HJ, Russell P, Aken B, Alfoldi J, Amemiya C, Azzouzi N, Baroiller JF, Barloy-Hubler F, Berlin A, Bloomquist R, Carleton KL, Conte MA, D'Cotta H, Eshel O, Gaffney L, Galibert F, Gante HF, Gnerre S, Greuter L, Guyon R, Haddad NS, Haerty W, Harris RM, Hofmann HA, Hourlier T, Hulata G, Jaffe DB, Lara M, Lee AP, MacCallum I, Mwaiko S, Nikaido M, Nishihara H, Ozouf-Costaz C, Penman DJ, Przybylski D, Rakotomanga M, Renn SC, Ribeiro FJ, Ron M, Salzburger W, Sanchez-Pulido L, Santos ME, Searle S, Sharpe T, Swofford R, Tan FJ, Williams L, Young S, Yin S, Okada N, Kocher TD, Miska EA, Lander ES, Venkatesh B, Fernald RD, Meyer A, Ponting CP, Streelman JT, Lindblad-Toh K, Seehausen O, Di Palma F. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513(7518):375–81.
5. Chan, Y. F., M. E. Marks, F. C. Jones, G. Villarreal, Jr., M. D. Shapiro, S. D. Brady, A. M. Southwick, D. M. Absher, J. Grimwood, J. Schmutz, R. M. Myers, D. Petrov, B. Jonsson, D. Schluter, M. A. Bell and D. M. Kingsley (2010). "Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer." Science 327(5963): 302–305.
6. Chen X, Zhang J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol. 2012;8(11):e1002784.
7. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, Kingsley DM. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. PLoS Biol. 2004;2(5):E109.
8. Conte MA, Kocher TD. An improved genome reference for the African cichlid, Metriaclima Zebra. BMC Genomics. 2015;16:724.
9. Dildrop R, Ruther U. Organization of Iroquois genes in fish. Dev Genes Evol. 2004;214(6):267–76.
10. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346(6213):1258096.
11. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–63.
12. Feijoo CG, Manzanares M, de la Calle-Mustienes E, Gomez-Skarmeta JL, Allende ML. The Irx gene family in zebrafish: genomic structure, evolution and initial characterization of irx5b. Dev Genes Evol. 2004;214(6):277–84.
13. Feiner N, Begemann G, Renz AJ, Meyer A, Kuraku S. The origin of bmp16, a novel Bmp2/4 relative, retained in teleost fish genomes. BMC Evol Biol. 2009;9:277.
14. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970;19(2):99–113.
15. Forslund K, Pekkari I, Sonnhammer EL. Domain architecture conservation in orthologs. BMC Bioinformatics. 2011;12:326.
16. Goldstein B, King N. The future of cell biology: emerging model organisms. Trends Cell Biol. 2016;26(11):818–24.
17. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(Database issue):D1079–85.
18. Hagen DW. Isolating mechanism in Threespine sticklebacks (Gasterosteus). J Fish Res Board Can. 1967;24(8):1637.
19. Hammock EA, Lim MM, Nair HP, Young LJ. Association of vasopressin 1a receptor levels with a regulatory microsatellite and behavior. Genes Brain Behav. 2005;4(5):289–301.
20. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014;157(6):1262–78.
21. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. The human phylome. Genome Biol. 2007;8(6):R109.
22. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of Phylogenomic data. Mol Biol Evol. 2016;33(6):1635–8.
23. Insel TR. The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior. Neuron. 2010;65(6):768–79.
24. Iwasaki K, Taguchi M, Bonkowsky JL, Kuwada JY. Expression of arginine vasotocin receptors in the developing zebrafish CNS. Gene Expr Patterns. 2013;13(8):335–42.
25. Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, P. Broad Institute Genome Sequencing, T. Whole Genome Assembly, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, E. S. Lander, F. Di Palma, K. Lindblad-Toh and D. M. Kingsley (2012). "The genomic basis of adaptive evolution in threespine sticklebacks." Nature 484(7392): 55–61.
26. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol. 2013;14(1):49–55.
27. Juntti SA, Hilliard AT, Kent KR, Kumar A, Nguyen A, Jimenez MA, Loveland JL, Mourrain P, Fernald RD. A neural basis for control of cichlid female reproductive behavior by prostaglandin F2alpha. Curr Biol. 2016;26(7):943–9.
28. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
30. Kocher TD. Adaptive evolution and explosive speciation: the cichlid fish model. Nat Rev Genet. 2004;5(4):288–98.
31. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38.
32. Kryuchkova-Mostacci N, Robinson-Rechavi M. Tissue-specificity of gene expression diverges slowly between Orthologs, and rapidly between paralogs. PLoS Comput Biol. 2016;12(12):e1005274.
33. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 2006;34(Database issue):D572–80.
34. Loh YH, Bezault E, Muenzel FM, Roberts RB, Swofford R, Barluenga M, Kidd CE, Howe AE, Di Palma F, Lindblad-Toh K, Hey J, Seehausen O, Salzburger W, Kocher TD, Streelman JT. Origins of shared genetic variation in African cichlids. Mol Biol Evol. 2013;30(4):906–17.
35. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.
36. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154(1):459–73.
37. Marques CL, Fernandez I, Viegas MN, Cox CJ, Martel P, Rosa J, Cancela ML, Laize V. Comparative analysis of zebrafish bone morphogenetic proteins 2, 4 and 16: molecular and evolutionary perspectives. Cell Mol Life Sci. 2016;73(4):841–57.
38. Martin A, Orgogozo V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. Evolution. 2013;67(5):1235–50.
39. McKinnon JS, Rundle HD. Speciation in nature: the threespine stickleback model systems. Trends Ecol Evol. 2002;17(10):480–8.
40. Mcphail JD. Predation and evolution of a stickleback (Gasterosteus). J Fish Res Board Can. 1969;26(12):3183.
41. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011;7(6):e1002073.
42. Nemudryi AA, Valetdinova KR, Medvedev SP, Zakian SM. TALEN and CRISPR/Cas genome editing systems: tools of discovery. Acta Nat. 2014;6(3):19–40.
43. Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Heriche JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R. TreeFam: 2008 update. Nucleic Acids Res. 2008;36(Database issue):D735–40.
44. Salazar VS, Gamer LW, Rosen V. BMP signalling in skeletal development, disease and repair. Nat Rev Endocrinol. 2016;12(4):203–21.
45. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. 2014;42(Database issue):D922–5.
46. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688–90.
47. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005;21(4):456–63.
48. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 2003;13(3):382–90.
49. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 2001;18(5):691–9.