Characterizing lineage-specific evolution and the processes driving genomic diversification in chordates

David E. Northover¹, Stephen D. Shank¹ and David A. Liberles¹²*

Abstract

Background: Understanding the origins of genome content has long been a goal of molecular evolution and comparative genomics. By examining genome evolution through the guise of lineage-specific evolution, it is possible to make inferences about the evolutionary events that have given rise to species-specific diversification. Here we characterize the evolutionary trends found in chordate species using The Adaptive Evolution Database (TAED). TAED is a database of phylogenetically indexed gene families designed to detect episodes of directional or diversifying selection across chordates. Gene families within the database have been assessed for lineage-specific estimates of dN/dS and have been reconciled to the chordate species to identify retained duplicates. Gene families have also been mapped to the functional pathways and amino acid changes which occurred on high dN/dS lineages have been mapped to protein structures.

Results: An analysis of this exhaustive database has enabled a characterization of the processes of lineage-specific diversification in chordates. A pathway level enrichment analysis of TAED determined that pathways most commonly found to have elevated rates of evolution included those involved in metabolism, immunity, and cell signaling. An analysis of protein fold presence on proteins, after normalizing for frequency in the database, found common folds such as Rossmann folds, Jelly Roll folds, and TIM barrels were overrepresented on proteins most likely to undergo directional selection. A set of gene families which experience increased numbers of duplications within short evolutionary times are associated with pathways involved in metabolism, olfactory reception, and signaling. An analysis of protein secondary structure indicated more relaxed constraint in β-sheets and stronger constraint on alpha Helices, amidst a general preference for substitutions at exposed sites. Lastly a detailed analysis of the ornithine decarboxylase gene family, a key enzyme in the pathway for polyamine synthesis, revealed lineage-specific evolution along the lineage leading to Cetacea through rapid sequence evolution in a duplicate gene with amino acid substitutions causing active site rearrangement.

Conclusion: Episodes of lineage-specific evolution are frequent throughout chordate species. Both duplication and directional selection have played large roles in the evolution of the phylum. TAED is a powerful tool for facilitating this understanding of lineage-specific evolution.

Keywords: Comparative genomics, Molecular evolution, Gene duplication, Pathway evolution, Protein structure

* Correspondence: daliberles@temple.edu

¹Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA
²Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA

© The Author(s). 2020 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background
As closely related species diverge after a speciation event, their genomes begin to accumulate changes that lead to molecular and phenotypic divergence. Speciation itself is a complex process in chordates that results from the gradual cessation of gene flow. As the isolated populations become separate species, mutations of different magnitudes affect the protein coding repertoire of the two diverging genomes. These changes include synonymous changes that only affect the nucleotide sites, nonsynonymous changes that affect the amino acid sites, and gene duplication and loss events, among other types of changes. A resource comparing chordate genomes in a phylogenetic context, The Adaptive Evolution Database (TAED) has recently been re-generated [33] extending previous versions that were released [46, 66].

The latest version of TAED contains gene families constructed systematically across chordate species as described in Hermansen et al. [33]. Gene families have been filtered for alignment quality and to prevent synonymous site saturation, with the oldest nodes in each rooted gene tree reflecting a speciation event of maximum age being the root of the chordate divergence. All pairwise alignments within each multiple sequence alignment had no more than 10% gaps and were at least 80% identical in non-gapped positions. This then created a trade-off between gene family ages (many had root nodes younger than the last common ancestor of chordates) and alignment quality, although homologous gene family relationships can still be identified through TAED. Gene families have been reconciled to the NCBI taxonomy [67] as a reference species tree and events of positive directional and diversifying selection detected using nonsynonymous to synonymous nucleotide substitution rate ratios in the branches model averaged across sites [83]. Gene families have also been used to identify duplication events using the SoftParsMap parsimony-based gene tree-species tree reconciliation software [9].

In addition to previous iterations of TAED, other studies have also sought to characterize the lineage-specific evolution of chordate genomes. This includes the generation of the Selectome Database [51] from Ensembl [2] data. Selectome extends gene family data automatically generated through the Ensembl pipeline which contains sequences from 68 different genomes. Gene families in Selectome are passed through stringent quality control steps following which tests of selection using branch-site models are implemented against tree topologies from Ensembl. While both Ensembl and Selectome examine evolution in a lineage-specific context, the method by which selection is detected varies, with Ensembl using pairwise analyses to calculate the normalized rate of nonsynonymous to synonymous substitutions (dN/dS) and Selectome using branch-site models of selection based on phylogenetic trees. Pairwise estimates of dN/dS do not account for phylogenetic information which limits the ability to understand evolution in a lineage-specific context, and prohibits detection of directional or diversifying selection on internal lineages. Branch-site models and branch models differ in their sensitivity (power) and selectivity (detection of false positives) [5, 25]. dS saturation is a potential problem for these approaches, with accuracy declining at dS ~ 3 [6].

Gene duplication is another important process to consider when assessing lineage-specific processes of evolution. As genes duplicate, they may undergo different evolutionary pressures and be either neofunctionalized, subfunctionalized, or pseudogenized [42]. In the classical model [55], duplicate gene copies can acquire mutations that lose (pseudogenize), change or gain (neofunctionalize) function mutations when the other copy retains the original function. Neofunctionalization, which can also occur to a gene subsequent to initial subfunctionalization, emerges as the dominant driver of evolution in duplicated genes in this model [35, 65]. As such it is one driver of lineage-specific differences in genome content. Subfunctionalization, the subdividing of functions from an ancestral state, can also lead to lineage-specific functional divergence of genes, without the gain of new functions in the genome as a whole. Without gene duplication as a source of genetic content unconstrained by negative selection, evolution tends to act in a conservative fashion [55].

TAED also presents a picture of lineage-specific evolution using pathway and structural information in addition to selection on individual protein encoding genes and gene duplication. Pathway level analyses of proteins may lead to understanding how proteins evolve in the context of a cell or organism, since proteins typically interact together in a pathway or network to achieve biological functions (phenotypes). Simulations have suggested that rate limiting steps are not evolutionarily stable over longer evolutionary periods [56, 57] and proteins currently involved in rate limiting steps may not remain so over long evolutionary periods. This suggests patterns that might be expected for gene-specific selective pressures in a pathway and how they relate to phenotypic evolution.

Two models for the evolution of pathways have been presented, the retrograde evolution model [34], proposing evolution to build a pathway backwards from the selected final product based upon affinity for related transition states at neighboring positions of a pathway and the patchwork model [38] suggesting that gene duplication retains catalytic mechanisms on widely distributed substrates that are dispersed throughout the network of pathways. A driver of mutational opportunity in both models is gene duplication. Analysis of protein
function can identify which model is best associated with the evolution of a given pathway, with evidence suggesting that the patchwork model is more common [48]. TAED compiles duplication and selection data compiled for pathways in a lineage-specific manner that can be viewed in this light.

Understanding the structural context of substitutions within a protein may elucidate the role of individual amino acid changes in potential functional shifts under positive selection, differentiating them from compensatory or stabilizing substitutions within the protein. Modeling the effects of amino acid substitutions can demonstrate changes in structure, dynamics, allosteric regulation, and ligand binding that can be used to identify functional shifts ([19]; see also [16]). Such modeling is limited however as the process is difficult and computationally intensive, with identification of fitness effects based upon biophysical models inexact. Measurements and models based on experimental work can also contribute to our understanding [14].

The structural context of mutations also impacts the substitution rate via negative selection. Requirements for folding stability drive lower substitutions in the protein core, while binding requirements on the ligand interface slow mutation as compared to the protein surface [28]. These constraints extend to functional requirements to avoid certain alternate states, including both selection against alternate folding states and substrates that result in deleterious interactions [47]. As protein structure diverges less observably than protein sequence over equivalent units of evolutionary time [36], similar structural constraints can be assumed to be approximately equivalently applicable to sequences diverged over relatively short evolutionary times.

Understanding how genes evolve and the processes by which they lead to novel adaptations in species is fundamental to understanding the genotype-phenotype map. Here we present some new characterizations of lineage-specific evolution utilizing the TAED database; we examine specific hypotheses across lineages, as well as characterizing processes at the levels of gene duplication, pathway evolution, and of protein structure.

**Results**

The Adaptive Evolution Database (TAED) contains ~3.2 million sequences from 3214 different chordate species. The database contains 143,806 individual genes families which are mapped to the chordate species tree. Twenty-three thousand nine hundred seventy gene families contained one or more branches with dN/dS > 1, indicating positive or directional selection acting on these lineages. When the dN/dS rates are high after controlling for dS saturation, the lineages are candidates for having undergone functional shifts. It is expected that the larger the dN/dS value for a given branch, the stronger the putative selective forces were to cause functional changes to the ancestral protein [73]. A list of the lineages with the largest dN/dS values where dS > 0.01 was generated, as these proteins constituted potential strong candidates for having undergone positive selection (Table 1). Of the top 30 lineages with the largest dN/dS values, values were found to range from 88.78 to 26.57. The families that these proteins come from are putatively involved in multiple different biological processes, many of which do not map to a KEGG pathway. Interestingly strong selection was found to have occurred on the branch leading from Boreoeutherian mammals in 9 of the top 30 instances of high dN/dS. This lineage constitutes species before the split of Laurasiatheria and Euarchontoglires, following the divergence of mammals. Additionally, strong selection was seen repeatedly on the lineage leading from Laurasiatheria which is the superorder containing cetaceans, carnivores, chiropterans, and ruminants. Functional shifts in these proteins may be responsible for some of the physiological and habitat differences between these groups and shared ancestors with carnivores and primates. Strong selection was seen to occur on the lineage leading from Neognathae which comprises most avian species. Pathways under selection along this lineage may indicate some of the functional differences between flightless birds which comprise the sister order Palaeognathae and other avians.

KEGG pathway mappings for the top 30 lineages with high dN/dS showed that selection may have acted on several different pathway types including metabolic pathway interactions, receptor signaling pathways, and immune response pathways. Selection can act directly on many different levels within an organism. It can occur at the DNA level, the protein level, the pathway level, and the phenotypic level. Understanding pathway evolution may ultimately be a better way to assess selection than current codon based methods [32].

**Enrichment analysis**

To gain a better understanding of pathways within TAED that are more common targets of directional selection, a test to determine which pathways were over or under represented for instances of putative positive selection was undertaken. Table 2 shows the list of the top 25 enriched KEGG pathways within TAED for directional selection. From the top 25 pathways that are over-represented in the database, 8 of the pathways are involved in metabolic reactions (the pathway labeled “Metabolic pathways” contains proteins from all metabolic pathways, and therefore is not a unique pathway). Metabolism, or the process of constructing useful cellular molecules, is essential for life. Given the vast array of different physiological and environmental conditions
Table 1 TAED gene family lineages with the largest dN/dS values where dS > 0.001

| Family dN/dS Value | Mapped Location on Chordate Species Tree (Start to End) | Family description | KEGG Pathways |
|-------------------|--------------------------------------------------------|--------------------|---------------|
| 151, 766          | Boreoeutheria to Laurasiathernia                       | splicing factor arginine/serine-rich 4/5/6 | Herpes simplex infection; Spliceosome |
| 30                | Cercopithecidae to Cercopithecida                       | transmembrane protein 91 isoform X1 |             |
| 60,787            | Euarchontoglires to Simiformes                         | fucose-1-phosphate guanylyltransferase | Metabolic pathways; Fructose and mannose metabolism; Amino sugar and nucleotide sugar metabolism |
| 23,133            | Aves to Aves                                           | galanin receptor 1 | Neuroactive ligand-receptor interaction |
| 296, 346          | Eutheria to Delphinida                                 | LOW QUALITY PROTEIN: probable N-acetyltransferase 16 |             |
| 21,900            | Boreoeutheria to Boreoeutheria                         | X-linked interleukin-1 receptor accessory protein-like 2 |             |
| 52,181            | Boreoeutheria to Laurasiathernia                       | unnamed protein product |             |
| 8186              | Sauria to Anolis carolinensis                         | protein Simiate |             |
| 9378              | Neognathae to Neognathae                               | palmitoyltransferase ZDHHC17 isoform X4 |             |
| 22,600            | Boreoeutheria to Laurasiathernia                       | potassium voltage-gated channel subfamily G member 3 isoform X1 |             |
| 4144              | Hystricognathi to Hystricognathi                       | pygopus homolog 1 |             |
| 14,875            | Laurasiathernia to Laurasiathernia                    | kinesin-like protein KIF16B |             |
| 12,213            | Camelidae to Camelus                                   | LOW QUALITY PROTEIN: dnaJ homolog subfamily B member 3 |             |
| 12,593            | Boreoeutheria to Boreoeutheria                         | tRNA (guanine(10)-N2)-methyltransferase homolog |             |
| 20,708            | Amniota to Amniota                                     | heparan sulfate 2-O-sulfotransferase HS2ST1 | Glycosaminoglycan biosynthesis - heparan sulfate / heparin |
| 32,532            | Boreoeutheria to Carlito syrichta                      | interleukin 8 | AGE-RAGE signaling pathway in diabetic complications; NOD-like receptor signaling pathway/Influenza A; Phospholipase D signaling pathway; Chemokine signaling pathway; Hepatitis B; Toll-like receptor signaling pathway; Legionelllosis; RIG-I-like receptor signaling pathway; Rheumatoid arthritis; Malaria; NF-kappa B signaling pathway; Shigellosis; Hepatitis CN on-alcoholic fatty liver disease (NAFLD); Pathways in cancer; Epithelial cell signaling in Helicobacter pylori infection; Amoebiasis; Bladder cancer; IL-17 signaling pathway; Chagas disease (American trypanosomiasis); Pertussis; Transcriptional misregulation in cancer; Salmonella infection; Cytokine-cytokine receptor interaction |
| 4273              | Laurasiathernia to Laurasiathernia                    | myelin protein zero | Cell adhesion molecules (CAMs) |
| 21,944            | Boreoeutheria to Boreoeutheria                         | adrenergic receptor alpha-2C | cGMP-PKG signaling pathway; Neuroactive ligand-receptor interaction |
| 14,303            | Neognathae to Neognathae                               | ATP-dependent DNA helicase PIF1 partial |             |
| 14,588            | Neognathae to Neognathae                               | organic solute transporter subunit alpha-like partial |             |
| 12,299            | Neognathae to Neognathae                               | phosphatidylinositol glycan class H | Glycosylphosphatidylinositol (GPI)-anchor biosynthesis; Metabolic pathways |
| 8762              | Myotis to Myotis                                       | doublesex- and mab-3-related transcription factor 2 isoform X1 |             |
| 55,196            | Murinae to Mus musculus                                | Isx protein |             |
that exist within chordate species, it is plausible that developing different metabolic strategies is a primary way for organisms to cope with their surroundings. As such, seeing that these pathways are often targets for directional selection is not surprising. Furthermore, it is evident from the list that pathways involved in immune response and cellular health have also been directly impacted by selection. Over-represented pathways involved in immune response included: Herpes simplex infection, Influenza A, Toxoplasmosis, and Th17 cell differentiation. It has been documented in the literature that selection against pathogens is a constant arms race that requires novel adaptations to overcome the constant pressures of pathogenic infection [15, 44, 78]; that these pathways should be over-represented for putative positive selection is not surprising. Additionally, pathways which alleviate physiological stress also appear to be over-represented for directional selection as seen in the pathways: fluid shear stress and atherosclerosis, non-alcoholic fatty liver disease, and chemical carcinogenesis. Cellular components were also found to be under selective pressure to evolve as seen in the pathways: protein processing in endoplasmic reticulum, RNA transport, lysosome, and peroxisome. Lastly, many lineages were found to have evolved under directional selection relating to olfactory transduction. Olfactory genes are the most duplicated genes within the human genome and are known to be largely expanded in other chordate species [54]. Olfactory sense is a primary means of communication, predation, and foraging for many species and thus is unsurprising that many lineages relating to this pathway have instances of dN/dS > 1.

Of the pathways found within TAED to be under-represented for functional shifts, surprisingly phototransduction was found to be included within the top 25 (Table 3). The ability to visually see pigments is important in both sexual selection and predation. In birds [12, 84], fish ([72, 74, 79];) and cetaceans [24] instances of positive selection have been discovered relating to selection on opsin and rhodopsin genes. Therefore, it is surprising that selection on this KEGG pathway would be under-represented within TAED. However, KEGG pathways for zeatin biosynthesis, penicillin and cephalosporin biosynthesis, bacterial secretion systems, and MAPK signaling pathway – plant, should be underrepresented in the database as these pathways are primarily involved in either plant or microbial systems and do not constitute meaningful pathways in chordates although orthologous proteins to some of the components to these pathways do exist in chordates, but may have different functions. RNA polymerase is a highly-conserved protein found throughout all domains of life, and therefore is unsurprising that the pathway for RNA polymerase would be under-represented for functional shifts within chordate species.

Another interesting question which was generated from structural elements contained in TAED was if some functional protein domains are more likely to experience elevated rates of evolution compared to others. To determine if this is true a systematic search was performed to determine what functional domain topologies are enriched within lineages in TAED that have signals for functional change (Table 4). Functional domains were annotated from the CATH database which assigns each domain a CATH classification. Annotations for this analysis looked at the topology level as it contains a wide array of functional domain annotations. The most over-represented domain/fold within TAED was the Rossmann fold which constituted approximately a quarter of all lineages in TAED with dN/dS > 1 that could map to a domain (the analysis did normalize for abundance in the database). The Rossmann fold is a common fold

| Family | dN/dS Value | Mapped Location on Chordate Species Tree (Start to End) | Family description | KEGG Pathways |
|--------|-------------|--------------------------------------------------------|-------------------|--------------|
| 14,119 | 31.0551     | Passeriformes to Passeriformes                         | cardiolipin synthase CMP-forming | Glycerophospholipid metabolism; Metabolic pathways |
| 39     | 30.43       | Boreoeutheria to Boreoeutheria                         | large subunit ribosomal protein L4e | Ribosome |
| 23,596 | 30.0168     | Neognathae to Neognathae                               | diacylglycerol cholinephosphotransferase | Phosphonate and phosphinate metabolism; Glycerophospholipid metabolism; Metabolic pathways; Choline metabolism in cancer; Ether lipid metabolism |
| 10,501 | 29.8433     | Homo sapiens to Homo sapiens                           | smoothelin isoform X5 | |
| 157, 103 | 27.6427    | Neognathae to Haliaeetus albicilla                      | nucleolar protein 7 partial | |
| 3717   | 27.4352     | Boreoeutheria to Boreoeutheria                         | glypican 4 | Wnt signaling pathway |
| 9725   | 26.5684     | Eutheria to Boreoeutheria                              | cilia- and flagella-associated protein 221-like partial | |

Table 1 TAED gene family lineages with the largest dN/dS values where dS > 0.001 (Continued)
comprised of a b-a-b-a-b (b – beta sheet, a – alpha helix) subunit motif and is commonly found within nucleotide-binding proteins [63]. Proteins that include this fold type include kinases, guanine nucleotide binding proteins (G proteins), proteins that bind cyclic adenosine monophosphate (cAMP), and NAD(P)-binding proteins [31]. These proteins are abundant within a cell and therefore proteins in which these domains reside are likely candidates for directional selection. However due to the nature and importance of nucleotide binding, it is unlikely that the Rossmann fold is under selection, but other domains within the same protein are as this domain is likely under strong negative constraint unless there are selective pressures on binding affinity or specificity. More structural analyses of the lineages under selection that contain the Rossmann fold would be warranted to examine this in more detail. The second most over represented domain topology was the Jelly Rolls fold which a subset of the beta-barrels superfamily. This fold type is composed of 8 beta-sheets which fold into a roll shape [1]. These folds are commonly found in viral capsid proteins [64]. It is possible that since these folds are commonly found in viral proteins that they evolve quickly and are prone to high mutation rates. This would suggest that protein families which contain this domain would be over-represented. The third most over-represented domain topology was TIM barrel folds. These are very common folds found with proteins that share alpha-beta structures. The TIM barrel folds are known to be highly promiscuous in sequence with many different sequences able to generate the TIM barrel fold. Therefore, there is biophysical flexibility for amino acids within these domains to be substituted while still maintaining the same domain structure [82]. These folds are

| Table 2 Pathways present in lineages under positive selection |
|---------------------------------------------------------------|
| KEGG Pathway                                               | Mapped Lineages Under Positive Selection | Lineages Under Positive Selection Mapped | Uncorrected P-value | FDR P-value | Bonferroni P-value |
| Metabolic pathways                                         | 7.63%                                   | 5.73%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Olfactory transduction                                     | 12.25%                                  | 2.67%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Biosynthesis of secondary metabolites                      | 7.85%                                   | 1.90%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Biosynthesis of antibiotics                                | 7.96%                                   | 1.11%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Neuroactive ligand-receptor interaction                    | 6.45%                                   | 1.05%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Microbial metabolism in diverse environments               | 7.86%                                   | 0.85%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Protein processing in endoplasmic reticulum                | 6.97%                                   | 0.72%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Purine metabolism                                          | 6.63%                                   | 0.69%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Herpes simplex infection                                   | 6.18%                                   | 0.68%                                     | 0.0047              | 0.0128      | 1.0000             |
| Carbon metabolism                                          | 8.63%                                   | 0.62%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| RNA transport                                              | 6.45%                                   | 0.59%                                     | < 0.0001            | < 0.0001    | 0.0056             |
| Influenza A                                                | 6.74%                                   | 0.59%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Fluid shear stress and atherosclerosis                     | 6.21%                                   | 0.54%                                     | 0.0061              | 0.016       | 1.0000             |
| Lysosome                                                   | 6.72%                                   | 0.52%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Glycerophospholipid metabolism                             | 7.61%                                   | 0.49%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Non-alcoholic fatty liver disease (NAFLD)                  | 6.19%                                   | 0.47%                                     | 0.0128              | 0.0325      | 1.0000             |
| Pancreatic secretion                                       | 8.86%                                   | 0.43%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Peroxisome                                                 | 7.37%                                   | 0.42%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Toxoplasmosis                                              | 6.50%                                   | 0.41%                                     | < 0.0001            | < 0.0001    | 0.0341             |
| Phosphatidylinositol signaling system                      | 6.70%                                   | 0.41%                                     | < 0.0001            | < 0.0001    | 0.0004             |
| Glycerolipid metabolism                                    | 10.21%                                  | 0.39%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Drug metabolism - cytochrome P450                           | 12.03%                                  | 0.39%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Th17 cell differentiation                                  | 6.38%                                   | 0.38%                                     | 0.0013              | 0.039       | 0.4944             |
| Valine, leucine and isoleucine degradation                 | 11.00%                                  | 0.38%                                     | < 0.0001            | < 0.0001    | < 0.0001           |
| Chemical carcinogenesis                                    | 9.32%                                   | 0.38%                                     | < 0.0001            | < 0.0001    | < 0.0001           |

The top 25 over-represented KEGG pathways with the highest % of lineages under positive selection mapping to a pathway. All pathways were significant at the 0.05 level after correction with the false discovery rate (FDR). Bold numbers indicate not significant at the 0.05 level. Lineages with dN/dS > 1 considered under positive selection.
in some cases known over longer evolutionary periods as folds that are structurally adaptable and evolve under relaxed selective constraint [17, 27, 45], consistent with their observation here in divergence among closely related species.

From the list of the top under-represented domain topologies (Table 5), two of the most under-represented domains were derived from the SMAD3 (mothers against decapentaplegic homolog 3) protein (smad3 chain A and Smad anchor for receptor activation chain B). The SMAD3 protein is involved in the signal trafficking of TGF-β which plays an important role in cell growth and death. This protein structure is known to contain two different domains, a DNA-binding domain and a protein-protein interacting domain. These two domains have been shown to be conserved across many species and play an essential role in the function of SMAD proteins [52, 53]. Accordingly, it is expected that these domains would be very limited in the rates at which they evolve and that they would evolve mostly under strong negative selection. Another interesting protein domain that was under-represented within the database was the fold for cAMP-dependent protein kinase. The primary enzyme which contains this domain is protein kinase A (PKA) which is involved in many different cellular pathways and plays a role in cell growth and differentiation, signaling, and migration [21]. As a central hub protein within a protein interaction network, it would be expected that this would be highly negatively constrained [58] and therefore domains that are essential to this protein are also under strong negative selection.
Duplication analysis

One important element of lineage-specific evolution is the expansion and contraction of genes within the genome. As genes duplicate they may undergo different evolutionary pressures and be either neofunctionalized, subfunctionalized, or pseudogenize [42]. Following the completion of the TAED database, it was interesting to determine if some gene families are more likely to undergo gene duplication events than others and what pathways these genes reside in. Are some pathways more flexible to gene duplication and dosage balance constraints [76] than others? A systematic examination of TAED gene family duplications was performed by scaling the number of duplication events detected within a family by the amount of time over which the family evolved. Three different proxies for time were used in the analysis, the maximum phylogenetic tree length measured in substitutions per site (Additional file 1: Figure S1), the median tree length measured in substitutions per site (Additional file 1: Figure S2), and the relative age of each family found by mapping the root of each gene tree to the chordate species tree (Fig. 1). Each analysis determined that there is a positive correlation between the number of duplications within the family and the amount of time over which the family evolved. Outliers from the regression line identified families that were highly duplicated over a shortened timespan. These families are also those with a high rate of duplication compared to other gene families. Table 6 shows the Cook’s distance calculations for the analysis using family
node age as a proxy for time and the corresponding gene families that were calculated to be furthest from the regression line. Cook’s distances for the maximum tree length and median tree length are found in Additional file 1: Tables S1 and S2, respectively. From the families with the largest Cook’s distance the number of times a highly duplicable family mapped to a give KEGG pathways was counted (Table 7). Pathway counts for the maximum tree length and median tree lengths were also calculated (Additional file 1: Tables S3 and S4).

The data shows metabolic pathways and olfactory receptors are consistently the top pathways where duplications occur. Olfactory receptors are known to be the largest expanded gene family [26], aligning our study with the currently known data.

Additionally, the top 25 most highly duplicable gene families included serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit epsilon isoform, abl interactor 1 - partial, aldolase B, guanine nucleotide-binding protein G(i) subunit alpha-1 - partial, and myosin regulatory light polypeptide 9. A further examination of the structural components and pathway components of these families may explain why they are more tolerable to duplication events and the mechanisms that are causing large gene family expansions. Interestingly, many of the most duplicated gene families mapped to KEGG pathways involved in immunity (HTLV-I infection;...
Herpes simplex infection; Epstein-Barr virus infection; Influenza A) and cancer (Pathways in cancer; Proteoglycans in cancer; Transcriptional misregulation in cancer; Viral carcinogenesis), possibly suggesting that duplication plays a strong role in this arms race.

Protein structure based analysis

The combination of gene families and information from the Protein Databank allows examination of how selection acts on a protein structural level. Gene families with associated protein structures were collated and aligned to the PDB alongside maximum likelihood ancestral sequences calculated by PAML.

The resulting profile is significantly different than the profile of non-substituted sites in the background on those lineages (Table 8). For both positively and negatively selected lineages, fewer substituted sites are buried relative to all sites on the protein; this is true both looking at all sites, and sites of any specific secondary structure, except for β-Sheet (p = 0.0361) and β-Bridge (p = 0.0081) sites on positively selected lineages, which was not significant after a multiple testing correction. The result in β-Bridge sites may simply be a matter of lower power due to the relatively small number of residues compared to most other secondary structures. β-Sheet sites are the most commonly substituted buried site on positive lineages (14.2744% vs 13.1684% for all helices), though α-Helix sites, as well as helices in general, are more common amongst all sites (15.9368 and 17.6017% vs 14.5822% for β-Sheet).

Negatively selected lineages consistently have an increase in the prevalence of exposed residues across all secondary structures, but this is not universal for positively selected lineages. α-Helix sites are the most frequent in the dataset and show no change in prevalence of exposed sites compared to non-substituted sites under positive selection. 310 Helix sites show an overall increase in substitution rates in negatively selected lineages, unlike other helixes but consistent with bends, turns and coil sites. This is likely linked to their lower stability and higher proportion of exposed vs buried sites.

In terms of secondary structure when both exposed and buried regions are considered together, substitutions are more likely to occur across less structured regions (Turns, Bends, and Coil areas) that are more likely to be exposed than buried on both positively and negatively selected lineages, but also β-Sheet sites on positively selected lineages and 310 Helix sites on negatively selected lineages. The changes in prevalence for each secondary structure is strongly related to the buried/exposed ratio of their own residues (particularly in negatively selected sites), so solvent exposure, while a significant factor, is not the only one. This corresponds with observations seen in other studies ([18] and studies cited therein).

The lack of significant change in β-Sheet buried sites on positively selected lineages, suggests that positive selection is freer to act on it than comparable α-Helix sites, which have a considerable drop in frequency amongst substituted (13.1684%) rather than all (17.6017%) sites. The β-Sheet site changes also point at differences between positive and negative selection. Unlike in positively selected lineages, in negatively selected lineages, a smaller proportion of substituted sites are buried β-Sheet sites compared to all sites. This suggests the difference on positively selected sites...
lineages is not simply due to lower fragility in $\beta$-Sheet structure, but an active role for $\beta$-Sheet internal structure in driving evolution of new functionality. It should also be considered that, in general, positively selected lineages have fewer $\alpha$-Helix (30.1108% vs 32.7617%) and more $\beta$-Sheet (21.7820% vs 19.8385%) sites compared to negatively selected lineages. Since, as discussed earlier, certain gene families and pathways are under more frequent positive selection than others, the lower selective constraint on $\beta$-Sheet sites has a long term impact on protein structure.

$\beta$-Bridge sites did not show a reduction in prevalence for substitutions on positively selected lineages. As these sites are used to hydrogen bond, particularly between $\beta$-sheets, the most likely source for these substitutions is to allow for protein restructuring. Purely compensatory driven changes are a less likely explanation, as negatively selected lineages where they are more likely than positively selected ones show a reduction in $\beta$-Bridge prevalence amongst substituted sites.

It should be noted that the same PDB structure is assumed to be applicable to all sequences in a gene family. As sequence pairs with divergence > 20% were split into separate families and as the median pairwise comparison among family members was 85% identity, the slow divergence of structural RMSD makes this a reasonable approximation [36]. Over longer evolutionary times [68, 69] and especially after lateral transfer events [60], repeated regions are known to lead to structural divergence.

### Gene family analysis of ornithine decarboxylase

Lastly TAED can be a valuable resource in understanding the lineage-specific evolution of individual gene families. To examine this, one gene family was selected based on criteria that it contained KEGG pathway mappings and structural information. The gene family that was analyzed encoded a putative ornithine decarboxylase. Ornithine decarboxylase is responsible for the decarboxylation of L-ornithine to putrescine. L-ornithine is

| Table 6 TAED gene families with many duplications based on family node age from summed branch lengths |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Cook’s Distance** | **Family Description** | **Family Age** | **Number of Duplications** |
| 0.0599 | serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit epsilon isoform | 684 | 729 |
| 0.0429 | receptor-type tyrosine-protein phosphatase F isoform X1 | 684 | 632 |
| 0.0364 | guanine nucleotide-binding protein G(i) subunit alpha-1, partial | 684 | 590 |
| 0.0358 | peptidyl-prolyl cis-trans isomerase A-like | 684 | 586 |
| 0.0338 | casein kinase I isoform gamma-2 | 684 | 572 |
| 0.0328 | transcription factor AP-2-alpha isoform X1 | 684 | 565 |
| 0.0310 | protein argonaute-3 | 684 | 552 |
| 0.0307 | serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B alpha isoform | 684 | 550 |
| 0.0303 | casein kinase I isoform epsilon | 684 | 547 |
| 0.0300 | cytoplasmic polyadenylation element-binding protein 4 isoform X1 | 684 | 545 |
| 0.0283 | late histone H2B.L4-like | 684 | 532 |
| 0.0266 | L-lactate dehydrogenase B chain-like | 615 | 566 |
| 0.0264 | mitogen-activated protein kinase 11 | 684 | 517 |
| 0.0262 | septin-14 | 684 | 516 |
| 0.0257 | serine/threonine-protein phosphatase 2B catalytic subunit beta isoform isoform X1 | 684 | 512 |
| 0.0247 | alpha-enolase | 684 | 504 |
| 0.0247 | mitogen-activated protein kinase 10 | 684 | 504 |
| 0.0240 | pre-B-cell leukemia transcription factor 1 | 684 | 498 |
| 0.0229 | heat shock protein HSP 90-beta-like | 684 | 489 |
| 0.0229 | potassium voltage-gated channel beta subunit | 684 | 489 |
| 0.0227 | polyadenylate-binding protein 1 | 615 | 529 |
| 0.0209 | protein yippee-like 2 | 684 | 471 |
| 0.0207 | sodium/potassium-transporting ATPase subunit alpha-4 isoform X1 | 684 | 469 |
| 0.0202 | myosin regulatory light polypeptide 9 | 684 | 465 |
| 0.0200 | aldolase B | 684 | 463 |
a key component to the urea cycle and the decarboxylation of L-ornithine signals the irreversible reaction of forming putrescine which is the first step in polyamine synthesis [59]. Polyamines are polycations able to bind negatively charged molecules such as DNA and RNA. Three primary polyamines are important regulators of the MAPK pathway which plays a role in cell proliferation: putrescine, spermidine, and spermine. Spermidine is produced from putrescine which can further impact apoptosis [50]. As these molecules play an important role in cell growth and cellular death, the committed step in the synthesis of polyamines would be hypothesized to be evolving under strong negative constraint.

An analysis of the TAED gene family showed six lineages with dN/dS > 1. These rates varied from a dN/dS rate of 2.0096 to 1.5451 (Table 9). Directional selection was found to have occurred on the lineage leading to Afrotherian mammals which are primarily localized to the continent of Africa and include: moles, elephants, manatees, and aardvarks. Other lineages with elevated rates of evolution were found for both Macaca mulatta (Rhesus macaque) and Dasypus novemcinctus (Nine-banded armadillo). Lastly, three different lineages involved cetacean species which may reflect the evolutionary pressures of moving from a terrestrial to an aquatic lifestyle. It was found that these instances of positive selection occurred following a duplication event, suggesting that the ornithine decarboxylase duplicate gene may have been under relaxed selective constraint following the duplication and not under the same strong constraints imposed by the polyamine synthesis pathway (Fig. 2). Although, since this protein was maintained and not lost over the 34 MYA of divergence between Orcinus orca (Killer whale) and Balaenoptera acutorostrata scammioni (Minke whale), it is likely that it has retained some functionality within these organisms.

To better understand the molecular mechanisms associated with the increased rate of evolution detected within the evolution of ornithine decarboxylase in cetaceans an examination of the ancestral changes mapped to the extant version of human ornithine decarboxylase was performed. For the changes on the branch Cetacea, it was seen that a nonsynonymous substitution occurred at site 238 with an asparagine substituting to an aspartic acid (N238D). This substitution is situated one residue from site 237 which is a known pyridoxal phosphate binding site [22] (Fig. 3). The decarboxylation of L-ornithine to putrescine is known to be a pyridoxal 5′-phosphate dependent reaction [37] and therefore changes to this site in the protein may impact the rate or ability to catalyze L-ornithine. The N238D substitution caused a substitution for an uncharged amino acid to be replaced by a negatively charged amino acid which could potentially impact the pyridoxal phosphate binding site (Fig. 3).

The active site of ornithine decarboxylase in humans is at residue 357 (Cystine - 357) [3]. While no substitutions were found at the active site, four different nonsynonymous substitutions were localized on the beta-sheets surrounding the active site. The substitutions P368Q, R375C, I376M, and R379H were all proximally close to the active site and may have been involved in remodeling of the active site for the cetacean duplicate of ornithine decarboxylase (Fig. 4). These mutations have impacted the ability of the protein in several ways, by either helping to stabilize the active site, change the specificity of the binding pocket, change the rate of the

---

**Table 7** TAED KEGG pathways based on duplication analysis using family node age from summed branch lengths

| KEGG Pathway                  | Number of mapping instances from highly duplicable families |
|-------------------------------|---------------------------------------------------------|
| Metabolic pathways            | 558                                                     |
| Olfactory transduction        | 198                                                     |
| Pathways in cancer            | 171                                                     |
| PI3K-Akt signaling pathway    | 130                                                     |
| Endocytosis                   | 130                                                     |
| HTLV-I infection              | 122                                                     |
| MAPK signaling pathway        | 121                                                     |
| Proteoglycans in cancer       | 106                                                     |
| Rap1 signaling pathway        | 98                                                      |
| Neuroactive ligand-receptor interaction | 96                                               |
| Ras signaling pathway         | 93                                                      |
| Regulation of actin cytoskeleton | 90                                          |
| Epstein-Barr virus infection  | 87                                                      |
| Purine metabolism             | 86                                                      |
| RNA transport                 | 85                                                      |
| Transcriptional misregulation in cancer | 83                                           |
| Protein processing in endoplasmic reticulum | 81                                           |
| Axon guidance                 | 80                                                      |
| mTOR signaling pathway        | 79                                                      |
| Focal adhesion                | 79                                                      |
| Viral carcinogenesis          | 78                                                      |
| Herpes simplex infection      | 77                                                      |
| cAMP signaling pathway        | 77                                                      |
| Ribosome                      | 75                                                      |
| Cytokine-cytokine receptor interaction | 72                                           |
reaction, or cause the active site to become inert. Further experimental validation would be necessary to understand how the N238D substitution and the putative remodeling of the active site may impact the function of the protein. However, evidence from TAED does suggest that cetacean ornithine decarboxylase has undergone functional shifts in several different sites which may impact the efficacy of the decarboxylation of L-ornithine to putrescine. Why this enzyme would be under selection within Cetaceans is also an unanswered question, but understanding the lineage-specific evolution of ornithine decarboxylase may help to decipher the mechanistic reasons for how cetaceans were able to re-adapt to life in the water.

Table 8 Sitewise substitution rates in TAED lineages sorted by selective pressure and structural features

| Substituted Selected Lineages (dN/dS > 1) | All Sites | Substituted Sites | Negatively Selected Lineages (dN/dS < 0.5) | All Sites |
|-----------------------------------------|----------|------------------|------------------------------------------|----------|
| Substituted Sites | p | All Sites | Substituted Sites | p | All Sites |
| Helix | 30.2826% | < 0.0001 | 34.0597% | 35.2377% | < 0.0001 | 36.6327% |
| Exposed | 17.1142% | 0.0002 | 16.4580% | 20.2511% | < 0.0001 | 17.3397% |
| Buried | 13.1684% | < 0.0001 | 17.6017% | 14.9866% | < 0.0001 | 19.2929% |
| α-Helix | 26.1346% | < 0.0001 | 30.1108% | 31.1229% | < 0.0001 | 32.7617% |
| Exposed | 14.3764% | 0.1659 | 17.6017% | 14.9866% | < 0.0001 | 15.1730% |
| Buried | 11.7582% | < 0.0001 | 15.9368% | 13.5979% | < 0.0001 | 17.5887% |
| 310 Helix | 3.4956% | 0.0098 | 3.3039% | 3.5115% | < 0.0001 | 3.2213% |
| Exposed | 2.3597% | < 0.0001 | 2.0045% | 2.4163% | < 0.0001 | 1.8907% |
| Buried | 1.1359% | 0.0004 | 1.2994% | 1.0952% | < 0.0001 | 1.3306% |
| n-Helix | 0.6524% | 0.8047 | 0.6449% | 0.6033% | < 0.0001 | 0.6497% |
| Exposed | 0.3780% | < 0.0001 | 0.2794% | 0.3098% | < 0.0001 | 0.2761% |
| Buried | 0.2743% | 0.0005 | 0.3655% | 0.2935% | < 0.0001 | 0.3736% |
| β-Sheet | 23.2104% | < 0.0001 | 21.7820% | 18.2981% | < 0.0001 | 19.8385% |
| Exposed | 8.9360% | < 0.0001 | 7.1998% | 7.3255% | < 0.0001 | 6.2661% |
| Buried | 14.2744% | 0.0361 | 14.5822% | 10.9726% | < 0.0001 | 13.5724% |
| β-Bridge | 1.1095% | 0.7913 | 1.0984% | 0.9888% | < 0.0001 | 1.0382% |
| Exposed | 0.5644% | 0.0004 | 0.4641% | 0.4876% | < 0.0001 | 0.4194% |
| Buried | 0.5451% | 0.0081 | 0.6343% | 0.5012% | < 0.0001 | 0.6188% |
| Turn | 12.0729% | < 0.0001 | 11.0540% | 12.3561% | < 0.0001 | 11.0859% |
| Exposed | 9.7554% | < 0.0001 | 8.3283% | 9.9588% | < 0.0001 | 8.1517% |
| Buried | 2.3175% | < 0.0001 | 2.7257% | 2.3973% | < 0.0001 | 2.9342% |
| Bend | 10.4763% | < 0.0001 | 9.7416% | 10.2628% | < 0.0001 | 9.6989% |
| Exposed | 7.9179% | < 0.0001 | 6.8552% | 7.8004% | < 0.0001 | 6.6547% |
| Buried | 2.5584% | < 0.0001 | 2.8864% | 2.4624% | < 0.0001 | 3.0443% |
| Coil | 22.4822% | 0.0006 | 22.2643% | 22.8565% | < 0.0001 | 21.7058% |
| Exposed | 15.2151% | < 0.0001 | 13.2976% | 15.3522% | < 0.0001 | 12.7858% |
| Buried | 7.6331% | < 0.0001 | 8.9667% | 7.5044% | < 0.0001 | 8.9201% |
| Buried (All Sites) | 40.4969% | < 0.0001 | 37.3970% | 38.8245% | < 0.0001 | 48.3826% |
| Exposed (All Sites) | 59.5031% | < 0.0001 | 52.6030% | 61.1755% | < 0.0001 | 51.6174% |

The distribution of substituted sites by secondary structure and solvent accessibility binned by the nature of selection are shown. Bolded items are significant (p < 0.00167 after multiple comparisons correction) based on parametric bootstrapping, n = 20000

Discussion
Understanding the mechanistic reasons that species diverge is of central importance to the field of molecular evolution. Gaining insight into how individual proteins evolve in context of the pathways in which they occur may help elucidate the underlying molecular mechanisms of speciation. Placing evolutionary events in the context of a species tree allows the interpretation of understanding how selective forces have varied across species. Here we have presented findings from The Adaptive Evolution Database (TAED) that have attempted to characterize the lineage-specific evolution of chordates. We know that selection can act on multiple levels within an organism, from the level of individual nucleotides to phenotypic
traits in a population. We therefore have examined the effects of directional selection at the domain level, gene level, and pathway level to better understand the dynamics of lineage-specific evolution. Examination of high level trends within TAED have confirmed that some pathways including those that are related to metabolism, immunity, and cell signaling have been repeated targets for functional change and may play important roles in species divergence. Additionally, we have shown that some protein families have undergone many duplication events which have impacted the evolutionary constraints of the duplicate pairs. These duplicated genes may evolve to new functions within the genome and develop new links within pathways. Tools developed on TAED can be utilized to find gene families that have undergone instances of adaptive evolution and help to propose hypotheses for how these genes have evolved.

Table 9 Lineages with dN/dS > 1 in Ornithine decarboxylase family

| dN/dS Value | Branch Start | Branch End | Mapped Branches |
|-------------|--------------|------------|-----------------|
| 2.0096      | Eutheria     | Afrotheria | Afrotheria       |
| 1.9244      | Dasypus novemcinctus | Dasypus novemcinctus | |
| 1.9712      | Cetacea      | Orcinus orca | Orcinus orca, Orcinus, Delphinidae, Odontoceti |
| 1.7717      | Cetacea      | Balaenoptera acutorostrata scammoni | Balaenoptera acutorostrata scammoni, Balaenoptera acutorostrata, Balaenoptera, Balaenopteridae, Mysticeti |
| 1.1272      | Cetacea      | Cetacea    | Cetacea         |
| 1.5451      | Macaca       | Macaca mulatta | Macaca mulatta |

Fig. 2 Gene tree for cetacean lineages of ornithine decarboxylase – Presented here is the gene tree taken from the TAED Tree Viewer for the TAED gene family 557. Lineages not associated with Cetaceans are collapsed. Internal nodes labeled with a while box are duplication events found within the tree. Nodes with solid grey dots represent speciation events. Nodes labeled in black indicate a leaf node. Lineages labeled in red have a dN/dS > 1 and the numbers along each branch are the associated dN/dS value for the given branch. Image was generated from the TAED Tree Viewer.
Fig. 3 Pyridoxal phosphate binding site for ornithine decarboxylase along the lineage of Cetacea – A protein homology model of the ancestral protein leading to Cetacea was created. Template for the model was from human ornithine decarboxylase (PDB:2OO0; chain A). Ancestral changes occurring on the lineage for Cetacea have been mapped to the model, sites colored in red indicate nonsynonymous changes in the ancestral protein, sites colored in dark grey are synonymous site changes. The site indicated in green is the pyridoxal phosphate binding site 238. The site adjacent to the binding site is the substitution N238D found on the ancestral lineage. Image was generated from Swiss-PdbViewer.

Fig. 4 Active site remodeling for ornithine decarboxylase along the lineage of Cetacea – A protein homology model of the ancestral protein leading to Cetacea was created. Template for the model was from human ornithine decarboxylase (PDB:2OO0; chain A). Ancestral changes occurring on the lineage for Cetacea have been mapped to the model, sites colored in red indicate nonsynonymous changes in the ancestral protein, sites colored in dark grey are synonymous site changes. The site indicated in gold is the active site cysteine-357. Remodeling of the active site can be seen in the changes P368Q, R375C, I376M, and R379H which are positioned around the loop containing the active site.
Not all parts of a protein are under the same selective constraints and residues located on the outside or surface of a protein may be more likely to evolve, and evolve at a different rate, than a residue which comprises the hydrophobic core of the protein. Our comparison of the solvent accessible surface area (SASA) and dN/dS showed that this holds for both positively selected and negatively selected lineages. It distinguishes differences between the action of the two kinds of selection beyond this by showing that while solvent accessibility is more exclusively the primary driver of changes in the nature of substituted sites on negatively selected lineages, positively selected lineages show relaxed selective constraint on β-Sheet and strengthen constraints on α-Helix sites.

Additionally, the relationship between the energetics of different substitutions and how they interplay with dN/dS could be explored by comparing dN/dS to the change in the change of free energy (ΔΔG) of a protein when different substitutions are introduced. Studies of this nature have examined how the thermodynamics of a protein influence the rate of dN/dS and how compensatory substitutions affect protein stability [61, 70]. Current evolutionary tests do not consider epistatic relationships within proteins, treating each site as acting independently from a statistical perspective.

Further, it is known that when Ne is large, selection is more efficient and the chance of an allele being lost from the population is small. However, when Ne is small the effects of genetic drift are greater and selection is less efficient [49, 75]. As such selection has limited ability to eliminate deleterious variants in chordates or fix advantageous changes, as chordate species have low effective populations sizes. Weber, et al. [80] found an unexpected negative correlation between Ne and dN/dS in bird populations, but found expected signals when considering the magnitude of biophysical effects of changes [80, 81].

TAED as a tool and resource in detecting episodes of lineage-specific evolution may also be useful in helping to understand the differences between directional selection and intra- and inter- molecular forces. Not all amino acid substitutions are the direct result of directional selection acting on a protein to functionally evolve. When physical changes within a molecule do occur, corresponding compensatory changes can occur which alleviate the deleterious effects of a mutation. These compensatory changes ensure that the newly substituted amino acid becomes the preferred amino acid for the residue in which it is located [61, 70]. Using traditional approaches of dN/dS it is difficult to differentiate between directional selection and compensatory changes as both aggregate across the branch. However, by examining changes in a lineage-specific context and determining when each substitution occurred along the lineage, it may be possible to begin to differentiate between these two processes.

The secondary structure analysis raises questions about the nature of the selective pressures on a protein-structure level, and points to the need for further investigation of β-sheet, α-helix, and 3_10 Helix structures and their role in protein evolution in particular.

**Conclusions**

TAED is a useful tool for understanding lineage-specific evolution and provides a source of data to develop further hypothesis-based inquiries into the mechanisms that drive diversification. In addition to providing an example of lineage-specific evolution in cetaceans, this work examined gene family evolution through the lenses of protein structure, co-evolution in pathways, as well as characterizing the duplication process within families. At the structural level, the study utilized the database to understand the differential patterns of amino acid substitution, including filtering by secondary structure, in comparing proteins under negative and positive selection. Overall, this work provides a further empirical window into the lineage-specific processes of evolution.

**Methods**

**Database construction**

The TAED database was constructed following the pipeline outlined in Hermansen et al. [33]. The pipeline includes generation of gene families from single-linkage clustering of BLAST results from chordate genes found on GenBank. A point accepted mutation (PAM) distance threshold of 120 was used for gene family construction. Gene families were refined for quality using an iterative method controlling for pairwise percent identity (> 80%) and the fraction of pairwise aligned gaps (< 10%). Gene families where then aligned using MAFFT [41] and phylogenetic trees were constructed using PhyML [30]. Gene tree – species tree reconciliation against the NCBI chordate taxonomy was implemented to determine putative duplication events and gene tree roots using Soft-ParsMap. Gene families were defined phylogenetically by the species tree except in cases where alignment quality prohibited this, as described here and in Hermansen et al. [33] (see [4] for a recent discussion of gene family construction methodology). Putative rates of evolution were then calculated using the branches model from PAML and dN/dS rates was computed. BLAST was then performed on TAED gene families against the KEGG database [40] to determine KEGG pathway relatedness and against PDB [10] to determine protein structure for each gene in TAED. All branches, including specifically those found to have a dN/dS > 1 (putatively evolving under positive selection) were mapped to the corresponding chordate species tree to determine along what lineage the elevated rates of evolution occurred and which proteins evolved rapidly on the same species tree.
lineage. Roots of all genes families were additionally mapped to the chordates species tree. To determine the approximate family root age for each gene family, information from TimeTree [43] was collected and root ages determined in MYA (millions of years ago). Domain classification information was gathered from the CATH database [71]. Putative functional annotations were assigned to each gene family based on NCBI nomenclature and KEGG pathway annotations when available.

Enrichment analysis
Over/under-represented KEGG pathway and domain analyses were performed with a BLAST search against the KEGG database of TAED gene families. KO numbers were assigned to each individual protein in TAED that contained a BLAST hit with an e-value <1e-10. This threshold was set so that all putative hits would be the result of orthologous descent instead of chance. The KO number from the top BLAST result was assigned to each TAED gene. KO numbers were then used to assess each putative biological pathway in which the protein is known to play a role. Over/under-representation of these pathways was then calculated using Fisher’s Exact test [23] and significance was estimated using an α-level of 0.05. The resulting p-values were corrected for multiple testing by performing a false discovery rate (FDR) analysis [8] with an FDR threshold of 0.05 and using a Bonferroni correction [13]. The FDR calculation was computed using the R statistical programming package [62]. A similar method was used to determine the over/under-representation of CATH domain topologies. The topology level classification was used as it represented a broad enough group that multiple topologies were found throughout TAED.

Duplication analysis
For each gene family in TAED, the root node of the family was mapped to its associated lineage on the chordate species tree. Nodes were then given approximate dates in MYA based on estimates from the TimeTree database [43]. The number of duplication events that occurred in each gene family was used as inferred by SoftParsMap [9] through reconciliation with the NCBI taxonomy for chordates. A linear regression was performed on the resulting comparison between the family root node ages and the number of duplication found within each gene family. The Pearson’s r coefficient was calculated for the resulting linear regression with a Pearson’s r = 0.59. Log scaled transformations of the data did not yield a strong regression coefficient.

Since families were sought that showed a high propensity for duplicability in a short amount of time, families that fell below the regression line were filtered out (Fig. 1). We also filtered out all families whose length was below the 5th percentile, since evolutionary forces may not have had time to act on families with so few substitutions. Outliers in the resulting set of families were detected using Cook’s distance [20], which measures the change in regression coefficients due to the removal of a data point, and is often used as a proxy for the influence of that point. Gene families were then sorted according to this distance (Table 6). Finally, the top quartile of families was measured using this distance and the number of times they occur in each KEGG pathway was counted (Table 7).

Additionally, to test how different proxies of time impacted the duplication analysis, two additional proxies for time were generated: the maximum tree length, and the median tree length. The maximum tree length estimated in substitutions per site was calculated for all gene tree topologies by taking the maximal tree length from root to leaf node for every TAED gene family as estimated by PhyML. The median tree length was calculated in a similar manner by taking the median of all distances between the root and leaf of the phylogenetic tree for each gene family. Additional file 1: Figures S1 and S2 illustrate the differences in the duplication distribution of the families based on the change of the time component to the analysis. Each axis was of the analysis was given the transformation $y = \log (1 + x)$ and the Pearson’s r coefficient was calculated. The resulting best coefficients for both the maximum tree length and the median tree were found when both axes were log-transformed. Cook’s distance was calculated for each proxy of time and the families with pathways from the families with the largest Cook’s distance to the regression line were tabulated.

Protein structure based analysis
Protein information was determined from stored PDB information associated with each gene family. To show that sites in different locations and belonging to different structures evolve at different rates, DSSP [39] values were used to ascertain the relative solvent accessibility (RSA) and secondary structure of individual sites within the protein was obtained. While newer and less approximate, but more computationally intensive methods than DSSP are available, a pilot analysis suggested that DSSP and more computationally intensive methods gave similar results for the purposes of this study. Membrane proteins and multimers were removed from the dataset based on identifying information in the PDB data. Sites were binned based on RSA using maximum surface areas from Tien et al. [77]; sites with a ratio greater than 0.20 were marked as exposed and buried otherwise, and then further categorized according to secondary structure. PAML analysis was used to determine the maximum likelihood ancestral sequence for each gene
associated with a protein and the results controlled for lineages with $\text{dN/dS} > 1$ and lineages with $\text{dN/dS} < 0.5$. $\text{dN/dS}$ values of 0 or between 0.5 and 1 were ignored, as were any sites that did not align with the PDB sequence or were not one of the most common 20 amino acids. To determine the significance of the calculated values, two-tailed non-parametric bootstrapping was performed. For each lineage, simulated datasets of size matching the total substituted residue count were generated, using the distribution of all sites on the respective lineages as a baseline.

**Gene family analysis of ornithine decarboxylase**

To demonstrate the application of lineage-specific analyses of evolution on specific gene families using TAED data, a gene family was selected for analysis based on the criteria that the gene family contained 3 or more lineages with $\text{dN/dS} > 1$ and it contained lineages that mapped to KEGG pathways and to a PDB structure. Using these criteria, the TAED gene family 554 (ornithine decarboxylase) was selected for further examination of lineage-specific evolution. $\text{dN/dS}$ estimates of each lineage were taken from the TAED database. A homology model was generated using Swiss-Model [11], with the automated build method. The top template used in the homology model was PDB entry 2OO0 chain A. Ancestral amino acids were mapped to the model. Active site and binding site information was taken from the PDB website for the same entry. Uniprot [7] data for ornithine decarboxylase was also used to make inferences into important catalytic sites within the molecule. Images of the homology model were generated using Swiss-PdbViewer [29].

**Supplementary information**

Supplementary information accompanies this paper at https://doi.org/10.1186/s12862-020-1585-y.

**References**

1. Aik W, McDonough MA, Thalhammer A, Chowdhury R, Schofield CJ. Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. Curr Opin Struct Biol. 2012;22(6):691–700. https://doi.org/10.1016/j.sbi.2012.10.001.
2. Atenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, Deluca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Py razor LP, Schreiber F, da Silva AS, Szklarczyk D, Train CM, Bork P, Lecompte O, von Mering C, Xenarios I, Sjölander K, Jensen LJ, Martin MJ, Miffato M, Quest for Orthologs consortium, Gabaldón T, Lewis SE, Thomas PD, Sonnhammer E, Dessimoz C. Standardized benchmarking in the quest for orthologs. Nat Methods. 2016;13(5):425–30. https://doi.org/10.1038/nmeth.3830.
3. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 2007;24(5): 1219–28. https://doi.org/10.1093/molbev/msm042.
4. Anisimova M, Liberis DA. Detecting and understanding natural selection. In: Codon evolution: Oxford University Press; 2012. p. 73–96. https://doi.org/10.1093/acprof:oso/9780199906153.003.0006.
5. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2004;32(000001):115–19. https://doi.org/10.1093/nar/gkh131.
6. Badger MJ, Khan NI, Weinberg RA, Menon UV. Ensembl 2017. Nucleic Acids Res. 2017;45(D1):D635–42. https://doi.org/10.1093/nar/gkw1104.
7. Altmann JR, Oliveira MA, Kern AD, Grishin NV, Phillips MA, Hackert ML. Crystal structure of human ornithine decarboxylase at 2.1 A resolution: structural insights to antizyme binding. J Mol Biol. 2000;295(1):7–16. https://doi.org/10.1006/jmbi.1999.3331.
8. Altmann JR, Oliveira MA, Kern AD, Grishin NV, Phillips MA, Hackert ML. Crystal structure of human ornithine decarboxylase at 2.1 A resolution: structural insights to antizyme binding. J Mol Biol. 2000;295(1):7–16. https://doi.org/10.1006/jmbi.1999.3331.
9. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 2007;24(5): 1219–28. https://doi.org/10.1093/molbev/msm042.
10. Anisimova M, Liberis DA. Detecting and understanding natural selection. In: Codon evolution: Oxford University Press; 2012. p. 73–96. https://doi.org/10.1093/acprof:oso/9780199906153.003.0006.
11. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2004;32(000001):115–19. https://doi.org/10.1093/nar/gkh131.
12. Badger MJ, Khan NI, Weinberg RA, Menon UV. Ensembl 2017. Nucleic Acids Res. 2017;45(D1):D635–42. https://doi.org/10.1093/nar/gkw1104.
53. Newfeld SJ, Wisotzkey RG. Molecular evolution of Smad proteins. In: Smad signal transduction. Dordrecht: Springer Netherlands; 2006. p. 15–35. https://doi.org/10.1007/1-4020-4799-6_1.

54. Nishima Y, Niki M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One. 2007;2(6):e607. https://doi.org/10.1371/journal.pone.0002078.

55. Ohno S. Evolution by Gene Duplication. Berlin: Springer-Verlag; 1970.

56. Orlenko A, Chi PB, Liberles DA. Characterizing the roles of changing population size and selection on the evolution of flux control in metabolic pathways. BMC Evol Biol. 2017;17(1):117. https://doi.org/10.1186/s12862-017-0962-7.

57. Orlenko A, Teufel AI, Chi PB, Liberles DA. Selection on metabolic pathway function in the presence of mutation-selection-drift balance leads to rate-limiting steps that are not evolutionarily stable. Biol Direct. 2016;11(1):31. https://doi.org/10.1186/s13062-016-0133-6.

58. Pang K, Cheng C, Xuan Z, Sheng H, Ma X. Understanding protein evolutionary rate by integrating gene co-expression with protein interactions. BMC Syst Biol. 2010;4:179. https://doi.org/10.1186/1752-0509-4-179.

59. Pegg AE. Regulation of ornithine decarboxylase. J Biol Chem. 2006;281(21):14529–32. https://doi.org/10.1074/jbc.R500031200.

60. Persi E, Wolf YI, Koonin EV. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. Nat Commun. 2016;7:13570. https://doi.org/10.1038/ncomms13570.

61. Pollock DD, Kinch LN, Liao Y, Grishin NV. Classification of proteins with shared motifs and internal repeats in the ECOD database. Protein Sci. 2016;25(7):1188–97. https://doi.org/10.1002/pro.2893.

62. R Core Team. R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing; 2016.

63. Rao ST, Rossmann MG. Comparison of super-secondary structures in Metazoa. Mol Biol Evol. 2016;33(12):3170–80. https://doi.org/10.1093/molbev/msv003.

64. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008;6(4):241–9. https://doi.org/10.1038/nrmicro1756.

65. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008;6(4):241–9. https://doi.org/10.1038/nrmicro1756.

66. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008;6(4):241–9. https://doi.org/10.1038/nrmicro1756.

67. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008;6(4):241–9. https://doi.org/10.1038/nrmicro1756.

68. Schaeffer RD, Kinch LN, Liao Y, Sheng H, Ma X. Understanding protein evolutionary rate by integrating gene co-expression with protein interactions. BMC Syst Biol. 2010;4:179. https://doi.org/10.1186/1752-0509-4-179.

69. Pegg AE. Regulation of ornithine decarboxylase. J Biol Chem. 2006;281(21):14529–32. https://doi.org/10.1074/jbc.R500031200.

70. Persi E, Wolf YI, Koonin EV. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. Nat Commun. 2016;7:13570. https://doi.org/10.1038/ncomms13570.

71. Orlenko A, Chi PB, Liberles DA. Characterizing the roles of changing population size and selection on the evolution of flux control in metabolic pathways. BMC Evol Biol. 2017;17(1):117. https://doi.org/10.1186/s12862-017-0962-7.

72. Orlenko A, Teufel AI, Chi PB, Liberles DA. Selection on metabolic pathway function in the presence of mutation-selection-drift balance leads to rate-limiting steps that are not evolutionarily stable. Biol Direct. 2016;11(1):31. https://doi.org/10.1186/s13062-016-0133-6.

73. Spielman SJ, Wilke CO. Maximum allowed solvent accessibilities of residues in proteins. PLoS One. 2013;8(11):e80635. https://doi.org/10.1371/journal.pone.0080635.

74. Van Valen L. Molecular evolution as predicted by natural selection. J Mol Evol. 1974;3(2):89–101 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6407466.

75. Weadick CJ, Loew ER, Rodd FH, Chang BSW. Visual pigment molecular evolution in the Trinidadian pike cichlid (Crenicichla frenata): a less colorful world for neotropical cichlids? Mol Biol Evol. 2012;29(10):3045–60. https://doi.org/10.1093/molbev/mss115.

76. Weber CC, Babolho B, Romiguer J, Ellegren H. K r / K c but not d N / d S correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. Genome Biol. 2014;15(12):542. https://doi.org/10.1186/s13059-014-0542-8.

77. Weber CC, Whelan S. Physicochemical amino acid properties better describe substitution rates in large populations. Mol Biol Evol. 2019. https://doi.org/10.1093/molbev/msz003 in press.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.