Whole-Genome Identification, Phylogeny, and Evolution of the Cytochrome P450 Family 2 (CYP2) Subfamilies in Birds

Daniela Almeida[1,2,†], Emanuel Maldonado[1,†], Imran Khan[1,2,†], Liliana Silva[1,2], M. Thomas P. Gilbert[3], Guojie Zhang[4,5], Erich D. Jarvis[6,7], Stephen J. O'Brien[8,9], Warren E. Johnson[10], and Agostinho Antunes[1,2,*]

[1]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Portugal

[2]Department of Biology, Faculty of Sciences, University of Porto, Portugal

[3]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark

[4]China National GeneBank, BGI-Shenzhen, Shenzen, China

[5]Centre for Social Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark

[6]Department of Neurobiology, Duke University Medical Center Durham

[7]Howard Hughes Medical Institute, Chevy Chase, Maryland

[8]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, Russia

[9]Oceanographic Center, Nova Southeastern University, Ft Lauderdale

[10]National Zoological Park, Smithsonian Conservation Biology Institute, Washington DC

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: aantunes@ciimar.up.pt.

Accepted: February 27, 2016

Abstract

The cytochrome P450 (CYP) superfamily defends organisms from endogenous and noxious environmental compounds, and thus is crucial for survival. However, beyond mammals the molecular evolution of CYP2 subfamilies is poorly understood. Here, we characterized the CYP2 family across 48 avian whole genomes representing all major extant bird clades. Overall, 12 CYP2 subfamilies were identified, including the first description of the CYP2F, CYP2G, and several CYP2AF genes in avian genomes. Some of the CYP2 genes previously described as being lineage-specific, such as CYP2K and CYP2W, are ubiquitous to all avian groups. Furthermore, we identified a large number of CYP2J copies, which have been associated previously with water reabsorption. We detected positive selection in the avian CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, and CYP2AC subfamilies. Moreover, we identified new substrate recognition sites (SRS0, SRS2_SRS3, and SRS3.1) and heme binding areas that influence CYP2 structure and function of functional importance as under significant positive selection. Some of the positively selected sites in avian CYP2D are located within the same SRS1 region that was previously linked with the metabolism of plant toxins. Additionally, we find that selective constraint variations in some avian CYP2 subfamilies are consistently associated with different feeding habits (CYP2H and CYP2J), habitats (CYP2D, CYP2H, CYP2J, and CYP2K), and migratory behaviors (CYP2D, CYP2H, and CYP2J). Overall, our findings indicate that there has been active enzyme site selection on CYP2 subfamilies and differential selection associated with different life history traits among birds.

Key words: avian genomes, cytochrome P450 (CYPs), substrate recognition sites (SRS), heme binding areas (HEM), positive selection.

Introduction

Cytochrome P450 (CYP) genes encode hemeproteins (Palmer and Reedijk 1991) in a wide variety of organisms (Nelson 2009). CYPs confer protection against reactive oxygen species that form in organisms after exposure to toxins and other environmental contaminants, including the drugs and carcinogenic compounds present in food. They are mainly expressed in the liver endoplasmic reticulum, but are also highly expressed in the small intestine and olfactory mucosa, suggesting that they have tissue-specific roles (Guengerich 2008). CYP enzymes are involved in phase I of detoxification (Konstandi et al. 2014) and are typically membrane-bound (Pochapsky et al. 2010). Usually they act as terminal oxidases in multicomponent electron-transfer chains called

© The Author(s) 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

P450-containing monooxygenase systems (Nebert and Gonzalez 1987) and contribute to the inactivation and excretion of several endogenous and exogenous noxious metabolites via urine or bile (Konstandi et al. 2014). The large variety and number of xenobiotics constantly encountered by species offer numerous challenges. To recognize and efficiently metabolize the array of common and novel substrates (Konstandi et al. 2014), CYPs have evolved multiple gene families consisting of several members with a diverse range of substrate specificities and regulation pathways (Nelson et al. 1993). These genes are suggested to be among the fastest-evolving gene systems (Konstandi et al. 2014) and have been linked with migratory behaviors, adaptations to novel habitats (Nelson et al. 1993; Nebert 2000; Nebert and Russell 2002; Nebert et al. 2013), more-efficient water retention (Jirimutu et al. 2012), and food selection (Sullivan et al. 2008). Their role in drug interactions and processing are also of interest to the pharmaceutical industry (Saxena et al. 2008). CYPs are vital in mechanisms of resistance to natural and synthetic compounds that potentially interfere with normal growth, development, and reproduction through their role in the processing of endogenous substrates (Danielson 2002). CYPs unique features, including high genetic diversity, broad substrate specificity, and catalytic versatility, enable them to deal with a wide variety of substrates (Scott and Wen 2001), fostering adaptation to and survival in new environments (e.g., migratory species, invasive species and resistance to pest controls drugs).

Since CYPs are a gene superfamily, their nomenclature system is based on a hierarchical clustering of genes into families and subfamilies (Nelson 2003). CYP families are named by number (e.g., CYP2), the subfamilies by capital letters (e.g., CYP2C), and the specific genes by a second number (e.g., CYP2C8). By convention, members of new CYP families must share greater than 40% amino acid identity, while members of subfamilies must share greater than 55% amino acid identity (Nelson 2003, 2009; Nelson et al. 2004).

In vertebrates, the CYP2 family (29 subfamilies) is one of the largest and most diverse and has the least-conserved nucleotide sequences (Nelson 1998, 2003). Typically, these genes have nine exons and are approximately 1,500 base pairs (bp) long. The CYP2U and CYP2R subfamilies are considered to be the most basal (Nelson 2003; Thomas 2007). In spite of their diverse and critical roles, understanding of the relationships among CYP2 subfamilies beyond the mammals is limited (Kirischian et al. 2011). Prior to our analyses, eight CYP2 subfamilies had been characterized (CYP2AB, CYP2AC, CYP2C, CYP2D, CYP2J, CYP2R, CYP2U, and CYP2W) in a limited number of bird species (chicken—*Gallus gallus*, turkey—*Meleagris gallopavo*, and zebra finch—*Taeniopygia guttata*) (Watanabe et al. 2013). Because avian species have diverse feeding habits, specific adaptations, and a worldwide distribution, they are exposed to a wide variety of compounds (environmental chemicals) and likely have developed an array of novel xenobiotic-metabolizing mechanisms (Watanabe et al. 2013). To test this idea, here we assessed the evolutionary history of the avian CYP2 family by conducting detailed analyses of gene content, adaptive evolution, and phylogenetic patterns across the whole genomes of 48 bird species from 36 orders from the recently conducted avian Phylogenomics Project (Jarvis et al. 2014; Zhang et al. 2014b), including species from the three major avian evolutionary groups: Palaeognathae, Galloanserae, and Neoaves.

Materials and Methods

CYP2 Gene Sequences

To characterize avian CYP2 genomics evolutionary diversity, we employed tBLASTn searches (Camacho et al. 2009) on 48 sequenced avian genomes from the Avian Phylogenomics Project in the GigaScience Database (Zhang et al. 2014a) and other sources (International Chicken Genome Sequencing 2004; Dalloul et al. 2010; Warren et al. 2010) using as query sequences individual CYP2 subfamily protein sequences annotated in Ensembl (Flicek et al. 2014) (release 75) for chicken (*G. gallus*), turkey (*M. gallopavo*), anole lizard (*Anolis carolinensis*), frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), and human (*Homo sapiens*). From the CYP2 sequences retrieved, only nucleotide sequences with more than 1,125 bp and high-identity (*e*-value $< 1e^{-5}$) were considered for further analyses. We then submitted these sequences from all 48 avian species to a BLASTx search (NCBI), which searches the protein database of all vertebrate species in NCBI using a translated nucleotide query optimized to find highly similar sequences, to accurately characterize the avian CYP2 subfamilies.

CYP2 Phylogenetic Analysis and Filtering of Gene Subfamily Data Sets

We performed a codon-based alignment of all identified avian nucleotide CYP2 sequences (genes and pseudogenes) along with some reference sequences of CYP2 subfamilies found in public databases (Ensembl release 75: http://www.ensembl.org and NCBI: http://www.ncbi.nlm.nih.gov, last accessed March 2014) for anole lizard, Chinese alligator (*Alligator sinensis*), green turtle (*Chelonia mydas*), frog, zebrafish, human, chicken, medium ground-finch (*Geospiza fortis*), common ostrich (*Struthio camelus australis*), and cormorant (*Phalacrocorax carbo*). Sequences were aligned in a "global data set" using MUSCLE (Edgar 2004) as integrated in the SEAVIEW 4.4.0 software package (Gouy et al. 2010). This alignment was tested for saturation bias using the Xia et al. statistic test (Xia et al. 2003) implemented in DAMBE 5.3.31 (Xia and Xie 2001).

To assess the adequacy of current consensus avian CYP2 subfamily nomenclature, a maximum likelihood (ML) phylogeny was estimated using our "global data set," which showed no significant evidence of saturation (P value < 0.05). This ML phylogeny assumed a General Time Reversible (GTR)

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

evolutionary model, with a proportion of invariable sites (I) and heterogeneity of substitution rates among sites modeled following a gamma distribution (G), as determined by jModelTest 2.1.1 (Darriba et al. 2012). The ML phylogeny was estimated using PHYML 3.0 (Guindon et al. 2010) with 100 bootstrap replicates and the Nearest Neighbor Interchange (NNI) branch search algorithm.

Based on this ML phylogeny, we retrieved the CYP2 nucleotide sequences from each well-defined clade in order to guarantee high identity among sequences that were compiled to create the "CYP2 subfamily data sets." Each subfamily data set was inspected closely to ensure there was only one CYP2 sequence per avian species and to remove pseudogenes and databases reference sequences. Several MUSCLE (Edgar 2004) alignments were constructed to corroborate previous avian CYP2 gene classifications, a process which led to the identification of smaller (incomplete) sequences, which were removed from their respective data sets. Avian CYP2 sequences with evidence of recombination or gene conversion events (Bonferroni corrected P values < 0.05) were also removed from the data sets. Recombination was assessed with the RDP4 software package using default settings and seven algorithms (RDP, GENECONV, Chimaera, MaxChi, SiScan, BootScan, and 3Seq) (Martin et al. 2010). Following this approach we obtained 17 "final avian CYP2 subfamily data sets": "CYP2AC," "CYP2AF," "CYP2C," "CYP2D," "CYP2H," "CYP2J," "CYP2J_1," "CYP2J_2," "CYP2J_3," "CYP2J_4," "CYP2J_5," "CYP2K_1," "CYP2K_2," "CYP2R," "CYP2U," "CYP2W_1," and "CYP2W_2," which were used for selection analyses. Supplementary table S1, Supplementary Material online, details which species are represented in each data set.

Ancestral Reconstruction Analysis of Avian CYP2 Subfamilies

To elucidate the evolutionary process of avian CYP2 subfamilies, we performed ancestral reconstructions using the COUNT software and employing default parameters (Csuros 2010). We used only the CYP2 subfamilies as identified above (threshold $>$ 1,125 bp), and the total evidence nucleotide species tree (TENT) of the Avian Phylogenomics Project (Jarvis et al. 2014), which was converted into an ultrametric format with the R8S 1.8 software following the author's instructions (Sanderson 2002). In this approach, the numerical gene profiles (number of genes present in each avian species per subfamily) were first converted into binary format (1—present or 0—absent) and the data were posteriorly analyzed using the Dollo parsimony model (Farris 1977).

Selection Analyses and Reassessment of the CYP2 Substrate Recognition Sites

We estimated the nature and strength of the evolutionary selection pressures at the molecular level by assessing ratios of nonsynonymous (dN) to synonymous (dS) substitution rates, or omega ($\omega = dN/dS$), where ω greater than, equal to and less than 1 is indicative of positive, neutral and negative selection, respectively. Strong negative selection ($\omega < 1$) pressures generally prevent the accumulation of amino acid changes in the regions of proteins that are essential for its structure and/or function (da Fonseca et al. 2007). In contrast, novel functionalities are often driven through positive selection ($\omega > 1$) favoring amino acid replacements in protein-coding genes (Antunes and Ramos 2007).

We started by employing site-models (Nielsen and Yang 1998; Yang et al. 2000) and branch-specific (Yang 1998; Yang and Nielsen 1998) likelihood analyses. For both analyses, we submitted the codon-based alignments of the "final avian CYP2 subfamily data sets" (supplementary table S1, Supplementary Material online) with the respective unrooted avian TENT (Jarvis et al. 2014) to the Codeml program from the PAML 4.7 package (Yang 1997, 2007).

We considered different codon substitution models (site-models) which allow the ω ratio to vary along sequences in different ways: 1) null models—M0 model that admits uniform selective pressure among sites and M1a, M7, and M8a models that do not allow sites with $\omega > 1$ and 2) alternative models—M3 model which assumes variable selective pressures among sites and M2a and M8 models which allow sites with $\omega > 1$ (Wong et al. 2004). In these analyses, likelihood-ratio tests (LRTs) were conducted by comparing the null models with the alternative models: M0 versus M3 (Anisimova et al. 2001, 2002; Suzuki and Nei 2001, 2002), M1a versus M2a (Nielsen and Yang 1998; Wong et al. 2004; Yang et al. 2005), M7 versus M8 (Yang et al. 2000), and M8a versus M8 (Swanson et al. 2003) to infer which models best fit the data. Whenever the LRT was significant (P value < 0.05) under the models M2a and/or M8, the codon sites under positive selection were identified using the Bayes Empirical Bayes (BEB) calculation, which analyzes the posterior probabilities (PP) for these sites (Yang et al. 2005). We only considered positively selected sites with PP $> 95\%$.

CYP2 subfamily data sets with evidence of positively selected sites (from site-models) were also submitted to branch-specific likelihood analyses (Yang 1998; Yang and Nielsen 1998) to assess if its ω-ratio varied significantly among distinct avian groups (branches of interest—foreground lineages) of the phylogeny. In these analyses, alternative branch models (with multiple ω-ratios for foreground lineages) were tested against simpler null models (which assume that all branches in the phylogeny are evolving at the same rate). The foreground lineages for alternative models were specified a priori based on the following categories: feeding habits (carnivorous, herbivorous, and/or omnivorous birds), habitat (dry, moist, and/or semi-moist) and migration (migratory and non-migratory). Supplementary table S2, Supplementary Material online contains the correspondence between each avian species and the above-mentioned traits. In order to guarantee a robust grouping of

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

branches into several partitions, where the strength of selection may be different (alternative models), we: 1) generated stochastic character maps for each trait across the previously obtained ultrametric TENT (Jarvis et al. 2014), following the method of Bollback (Bollback 2006), as implemented in *phytools* (Revell 2012) and *geiger* (Pennell et al. 2014) R packages, using R 3.2.2 software (R Core Team 2015); 2) labeled the unrooted TENT according to these mapping results; and 3) trimmed the resulting labeled TENT to retain only the avian species represented in each one of the corresponding "final avian CYP2 subfamily data set" (supplementary figs. S1–S11, Supplementary Material online).

For both site-models and branch-specific selection analyses we applied the F3x4 codon model (Yang and Nielsen 2008) allowing for ML estimation of κ (transition/transversion ratio) and ω. All the models were run several times, adjusting the initial κ and ω values in order to avoid possible local-likelihood peaks. For all model comparisons, the hypothesis decision threshold was calculated by doubling the difference between the alternative and null model log likelihood ($2\Delta lnL$) and assuming that the null distribution of these results could be approximated by a chi-square (χ^2) distribution (P value < 0.05). The number of degrees of freedom (df) was calculated as the difference in the number of estimated parameters between the models (Yang 2000; Wong et al. 2004). We used the IMPACT_S software to automate these calculations (Maldonado et al. 2014).

To search for site-specific amino acid properties that are being preserved (conserved properties) or modified (changing properties) through the evolutionary process, we used the PRoperty Informed Models of Evolution (PRIME) method (Pond, unpublished work) as implemented in the Datamonkey webserver (Pond and Frost 2005; Delport et al. 2010). The PRIME method considers two predefined sets of five physicochemical amino acid properties. These include five empirically measured amino acid properties proposed by Conant et al. (2007): 1) chemical composition (McClellan et al. 2005) of the side chain [CC], 2) residue polarity [P], 3) volume [V] of the residue side chain, 4) isoelectric point [pHi] of the side chain, and 5) hydropathy [H] (Conant et al. 2007) and five composite properties proposed by Atchley et al. (2005): 1) polarity index [P], 2) secondary structure factor [SS] (McClellan et al. 2005), 3) volume [V], 4) refractivity [μ], and 5) isoelectric point [pHi] (Atchley et al. 2005). The estimates of amino acid exchangeabilities implemented by this method are based on multiple tests performed on the same residue site. Therefore this method includes the Bonferroni correction to control the number of false positives reported at a site. From the sites reported by this approach, we only considered and analyzed those that were coincident with previously identified sites with significant evidence of positive selection by BEB.

CYP2 enzymes have substrate recognition sites (SRS), where the amino acids are close to the ligands and thus influence substrate recognition and/or binding (Gotoh 1992) and induce chemical and structural variations that are reflected on the size, shape, and chemical features of substrates and products (da Fonseca et al. 2007). To map the positively selected sites onto the tridimensional (3D) structure of the CYP2 subfamilies and to facilitate visualization of sites hypothesized to be under important SRS, we first considered the available information about six CYP2 SRS (Gotoh 1992). These six SRS were described by Gotoh in 1992 (SRS*) from the alignment of mammalian CYP (1, 2 and 3) sequences with the CYP101A1 sequence from the bacterium *Pseudomonas putida*, whose substrate-binding sites were identified by X-ray crystallography of a substrate-bound form (Poulos et al. 1987). To update these six regions, we first searched the Protein Data Bank (PDB) database (release April 8, 2014) to obtain CYP2 X-ray crystal structures with the appropriate ligand annotations, to further perform their amino acid alignment (MUSCLE) (Edgar 2004) with a consensus sequence showing the six Gotoh's SRS. Through consensus alignments—using the GENEIOUS 5.6.7 *consensus align* option (Kearse et al. 2012)—of the sequence containing the annotations of updated SRS with each one of the avian CYP2 subfamily alignments, we verified if their positively selected sites were within important SRS. We then performed homology modeling of the 3D structure of the avian CYP2, which showed evidence of positive selection, using the SWISS-MODEL webserver (Arnold et al. 2006; Biasini et al. 2014). If the avian predicted models were not reliable, we only mapped the positively selected sites onto the 3D structure when such models (from the same subfamily) were available in the PDB database for other vertebrates (i.e., human CYP2C and CYP2D PDB codes: 2VN0 and 3TDA, respectively). The superimposition, visualization and manipulation of the 3D structures were performed with PYMOL 1.5.0.4 software (DeLano 2002).

Statistical Analyses on Trait Associations

In order to understand if the distribution of the number of CYP2 genes could be used to differentiate bird species according to their migratory/nonmigratory behavior, we performed a linear discriminant analysis (LDA). The classification variable was the migratory behavior of the species, with two classes (migratory/non-migratory) and the independent variables were the ten CYP2 subfamilies (CYP2F and CYP2G were excluded as they are outliers). The percent of correct predictions in each class was evaluated by cross-validation. These analyses were performed with the *lda* function from *MASS* R package, using R 3.2.2 software (R Core Team 2015).

Results

CYP2 Genes Have Diverse Paralogs Depending on Subfamily

The BLAST analyses performed in the 48 avian genomes identified 642 CYP2 (including genes and pseudogenes) with

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

sequence identity with other gene members available varying between 61% and 100%. Following the current nomenclature system (Nelson 2003, 2009; Nelson et al. 2004), these BLAST results identified 12 CYP2 subfamilies in birds: CYP2AC, CYP2AF, CYP2C, CYP2D, CYP2F, CYP2G, CYP2H, CYP2J, CYP2K, CYP2R, CYP2U, and CYP2W (fig. 1).

Subfamilies CYP2C (38 genes and 4 pseudogenes across 40 species), CYP2D (36 genes and 1 pseudogene across 37 species), CYP2H (40 genes and 1 pseudogene across 39 species), CYP2J (205 genes and 13 pseudogenes across 48 species), CYP2K (84 genes and 3 pseudogenes across 47 species), CYP2R (43 genes across 43 species), CYP2U (43 genes across 43 species), and CYP2W (77 genes and 2 pseudogenes across 46 species) were widely present in birds (fig. 1). We also found members of the CYP2AC subfamily in birds (22 genes and 2 pseudogenes across 24 species). Our analyses also revealed several previously undescribed subfamilies in birds: CYP2F in the grey-crowned crane genome, CYP2G in the chimney swift, and 26 CYP2AF genes in 26 avian species (fig. 1). Overall, most subfamilies only had one paralog across species (CYP2C, 2D, 2H, 2R, 2U, 2AC, and 2AF), but the CYP2K and CYP2W had two paralogs each, whereas the CYP2J had between 1 and 7 paralogs per species (fig. 1).

The ML phylogeny from the "global alignment" replicated the currently accepted basal nomenclature (based on BLAST analyses) with high node bootstrap support (73–100%) for each CYP2 subfamily clade (fig. 2). However, the support within each subfamily across species were not as clearly resolved (< 50% node bootstrap support). This is consistent with the findings of (Jarvis et al. 2014), where most individual gene trees of birds do not have enough phylogenetic resolution or have a large amount of incomplete lineage sorting such that no gene tree completely matches the genome-scale species tree. Therefore, our analyses of the avian CYP2 subfamily evolution and selection analyses were done using the more-robust genome-scale TENT tree as reference (Jarvis et al. 2014).

Evolutionary Process of CYP2 Subfamilies among Avian Lineages

The ancestral reconstruction (fig. 3), performed by COUNT software (Csuros 2010), suggested that: 1) the most recent common ancestor of modern birds must have had elements of CYP2C, 2J, 2K, 2R, 2U, 2W, 2AC, and 2AF and then, over time, several genes might have been lost, mainly during the evolution of Neoaves; 2) CYP2D and CYP2H subfamilies are likely to have been lost in the Paleognathae lineage. Moreover, we detected a very large number of CYP2AC and CYP2AF subfamily genes that were lost, in sharp contrast with CYP2J, which is the most conserved subfamily across birds. Interestingly, the emperor penguin lost seven subfamilies during its evolution, in contrast with its close relative, the Adelie penguin, which lost only one (fig. 3).

Active Sites of CYP2 Enzymes Have Been Differentially Selected

The PAML LRT site-model analyses of the filtered 17 final CYP2 data sets' without pseudogenes, revealed significant evidence of positive selection in six of the 12 avian CYP2 subfamilies: 1) CYP2C, 2) CYP2D, 3) CYP2H, 4) "CYP2J" ("CYP2J_1," "CYP2J_2," "CYP2J_3," "CYP2J_4," and "CYP2J_5" data sets), 5) CYP2K ("CYP2K_1" and "CYP2K_2" data sets), and 6) CYP2AC. To determine the impact of positive selection at these sites, we performed a detailed molecular analysis based on the 3D structure of the CYP2 proteins, including SRS* sites identified by Gotoh (1992). Search of the PDB database (release April 8, 2014) identified 61 available 3D CYP2 (A, B, C, D, E, and R subfamilies) structures and their sequences. Our comparative analyses (supplementary fig. S12, Supplementary Material online) of these PDB sequences and their respective ligand annotations with a consensus sequence with the six SRS* revealed several SRS areas (updated SRS): SRS0 (new), SRS1, SRS2_SRS3 (new, resulting from the fusion of SRS2* and SRS3*), SRS3.1 (new, between the SRS3* and SRS4*), SRS4, SRS5, and SRS6 (fig. 4). Comparative analyses of updated SRS with each one of the avian CYP2 subfamily alignments allowed the identification of several positively selected sites within important SRS. Below we highlight the SRS sites for each of six subfamilies with evidence of positive selection (the numbering of the sites is based on their corresponding amino acid sequences shown in supplementary table S5, Supplementary Material online).

CYP2C: Model M2a indicated that approximately 2% of the sites were under positive selection ($\omega 2 = 4.159$) whereas model M8 showed that approximately 3% were under positive selection ($\omega = 3.240$) (supplementary table S3, Supplementary Material online). These included the following seven sites: 239, 254, 281, 333, 369, 379, and 453 (supplementary table S4, Supplementary Material online). Three of these are located within SRS: sites 239 and 254 are located within SRS2_SRS3 (fig. 5A) and site 369 is within SRS5 (fig. 5B), a recognized heme binding area (HEM). The PRIME analyses suggested that several of the positively selected sites detected by PAML (Yang 2007) would have amino acid changing properties that could affect the chemical composition, secondary structure, isoelectric point and refractivity of the CYP2C sequences (supplementary table S4, Supplementary Material online). This is the case of sites 239 and 281 (fig. 5A), 369 (fig. 5B), and 453 (fig. 5C) (supplementary table S4, Supplementary Material online).

CYP2D: Model M2a indicated that approximately 5% of the sites were under positive selection ($\omega 2 = 2.983$) whereas model M8 indicated 8% ($\omega = 2.504$) (supplementary table S6, Supplementary Material online). Of the seven sites that were identified to be under positive selection by both methods (54, 74, 123, 236, 240, 359, and 437— supplementary table S5, Supplementary Material online, line 2), two (54 and 74) were

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

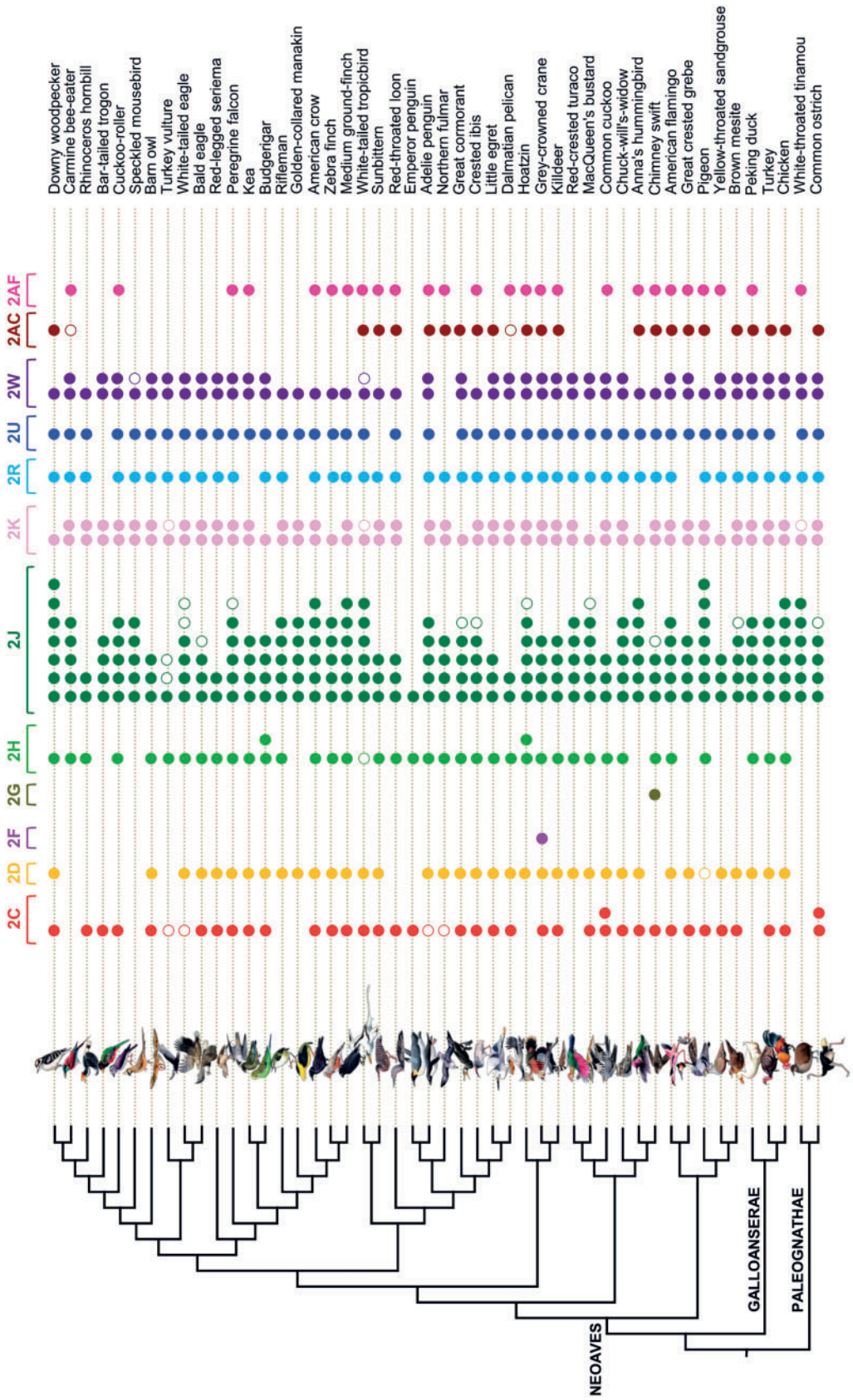


Fig. 1.—Identification of the CYP2 genes found in 48 avian genomes. At the bottom is shown a representation of the avian TENT tree (Jarvis et al. 2014). Middle is full and open circles symbolizing the genes and pseudogenes, respectively. Left row list the ends of the CYP2 gene subfamily names. Top row list the avian species common names.

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

Fig. 2.—Evolutionary relationships of avian CYP2 subfamilies. The phylogenetic tree was built in PHYML 3.0 software using the ML method, with 100 bootstrap replicates and the NNI branch search algorithm. The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the 642 avian CYP2 nucleotide sequences from 48 avian genomes and 31 CYP2 nucleotide sequences from reference species available in the public databases. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test is shown next to the branches or represented by an asterisk mark (* indicating 100% support) for each one of the CYP2 subfamily clades. Values less than 50% support are not shown.

located within the SRS1 and HEM regions (fig. 6), which have been linked with CYP2D catalytic activity. The PRIME results showed that some of the positively selected sites located outside the active site areas (that were also detected by PAML analyses) also would likely change the properties of the selected amino acid and thus affect the hydropathy (123), polarity (236), and volume (359) of this enzyme (supplementary table S7, Supplementary Material online).

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

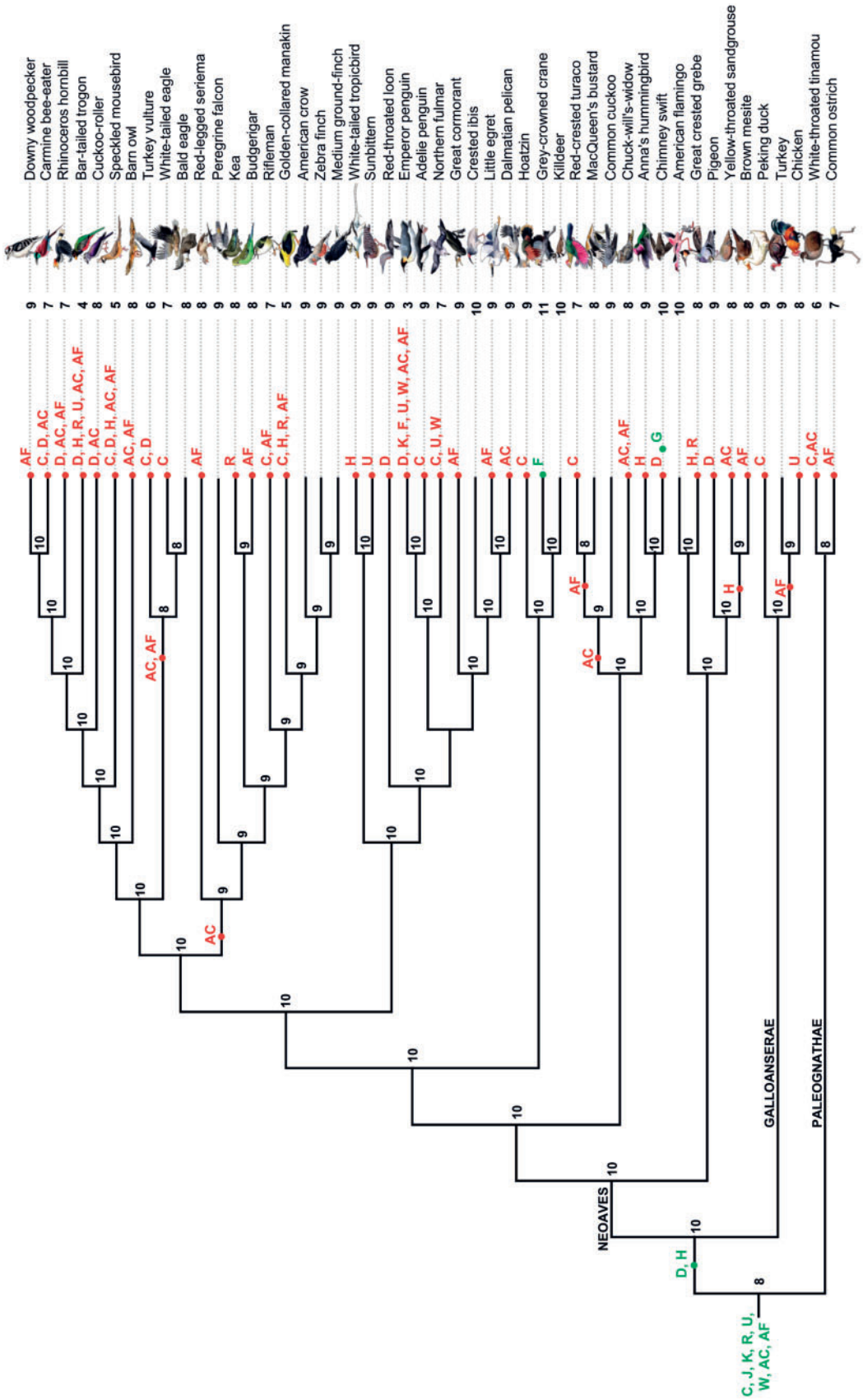
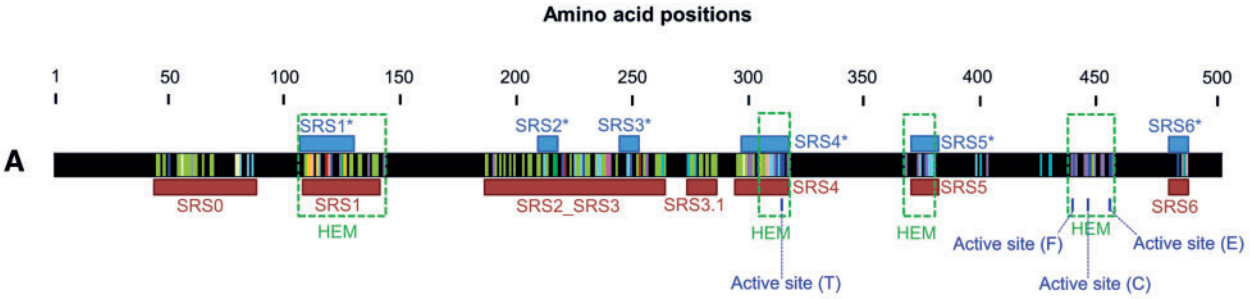


Fig. 3.—Evolutionary process of avian CYP2 subfamilies. Green circles indicate the gain of the corresponding CYP2 subfamilies while red circles indicate their loss. Numbers along the phylogenetic tree denote the number of CYP2 subfamilies present in each moment of the avian evolution.

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024



B

Binding regions	Start position	End position	Length
SRS0	45	86	42
SRS1	109	141	33
SRS1*	108	129	22
SRS2*	210	217	8
SRS2_SRS3	187	264	78
SRS3*	245	252	8
SRS3.1	274	286	13
SRS4	295	317	23
SRS4*	298	316	19
SRS5	371	381	11
SRS5*	371	381	11
SRS6	482	489	8
SRS6*	482	489	8
HEM	439	457	19
HEM	109	143	35
HEM	306	317	12
HEM	368	380	13
Active site (C)	447	447	1
Active site (E)	456	456	1
Active site (F)	440	440	1
Active site (T)	313	313	1

Fig. 4.—SRS of CYP2 proteins. The SRS are inferred from comparative analyses of amino acid sequences from available 3D CYP2 structures (supplementary fig. S12, Supplementary Material online) with the six SRS previously identified by Gotoh (1992). (*A*) Schematic representation of the consensus sequence resulting from the alignment of 61 CYP2 amino acid sequences (supplementary fig. S12, Supplementary Material online), containing ligand annotations, with the annotations of the six SRS (in blue: SRS1*, SRS2*, SRS3*, SRS4*, SRS5*, and SRS6*) inferred by Gotoh (1992). The sites where these substrates bind are represented by several colors, according with the number of different ligands. Red boxes represent new defined SRS, based on the available information about CYP2 ligands interacting with the 3D structure of CYP2. The high rate of binding sites, distributed by well-defined regions, allowed us to define seven distinct regions that were named SRS0, SRS1, SRS2_SRS3 (resulting from the fusion of the Gotoh's SRS2 and SRS3), SRS3.1, SRS4, SRS5, and SRS6 (all in red), in order to keep the nomenclature previously used by the referred author. The heme binding regions are denoted by HEM. (*B*) Identification of the amino acid boundaries and length of the referred SRS.

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

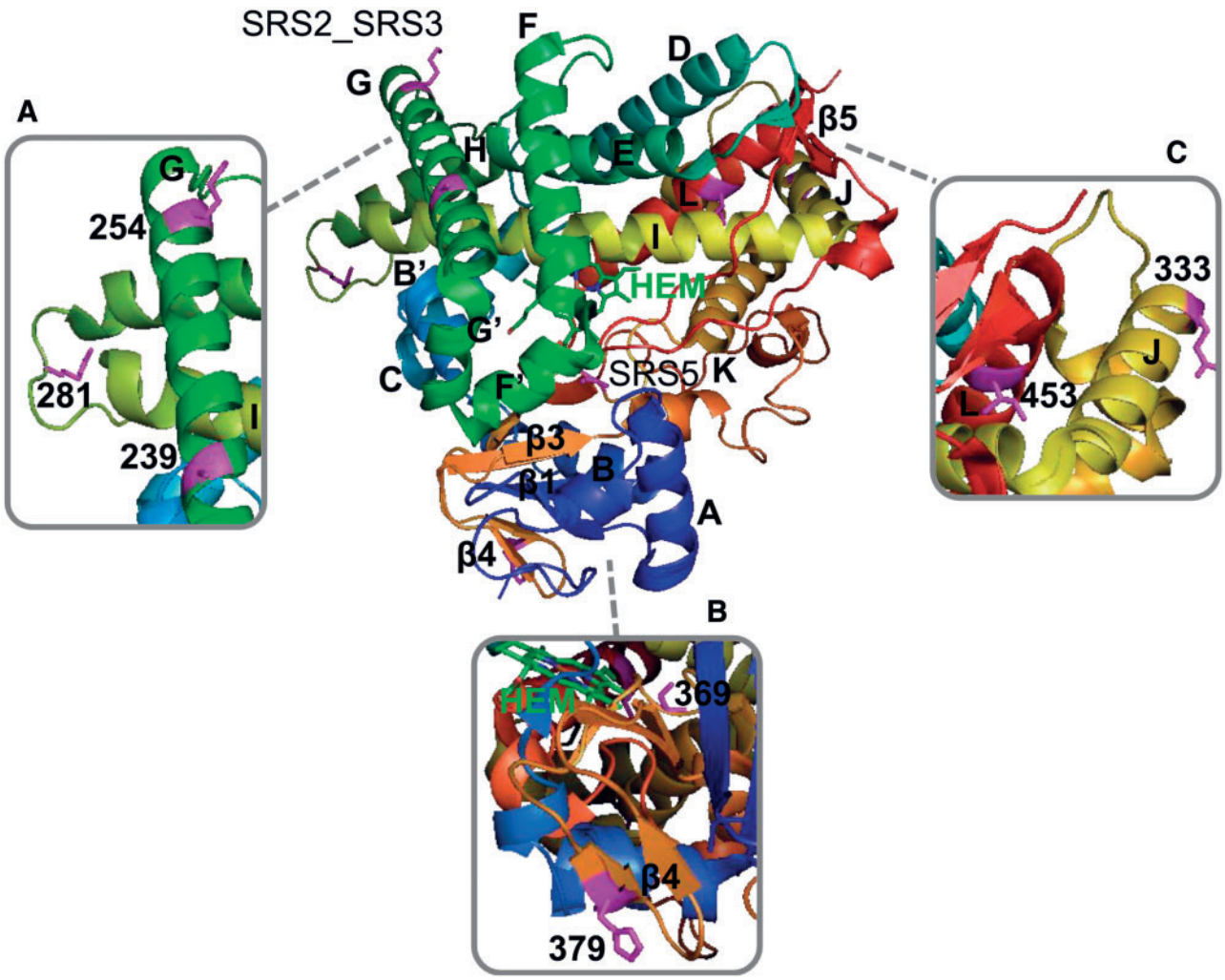


Fig. 5.—3D analyses of sites detected to be under positive selection in the avian CYP2C. The CYP2C 3D structure predicted in this study is superimposed to its CYP2C8 template (2VN0 human). The positively selected sites are shown as pink sticks, indicating the corresponding amino acid. The regions corresponding to the helices are named with the corresponding capital letter and the β-structures are named with a β followed by a number. HEM represents the heme group. Two SRS regions are represented: SRS2_SRS3 (top) and SRS5 (bottom).

CYP2H: Model M2a indicated that approximately 12% of the sites were under positive selection ($\omega 2 = 2.503$) whereas model M8 indicated 16% ($\omega = 2.191$) (supplementary table S8, Supplementary Material online). The BEB analysis from both models identified 21 positively selected sites (supplementary table S9, Supplementary Material online). Sites 38, 45, and 71 are located within the newly defined SRS0 (supplementary table S9, Supplementary Material online) and site 102 is within the SRS1 and HEM regions (supplementary table S9, Supplementary Material online). Sites 212, 227, 228, and 248 were located within the newly determined SRS2_SRS3 region (supplementary table S9, Supplementary Material online) and sites 236 and 240 were in SRS2_SRS3 which matches SRS3* (supplementary table S9, Supplementary Material online and fig. 4). Site 305 corresponds with the SRS4 and HEM binding site. Site 365 is located in the HEM binding region and also in SRS5 (supplementary table S9, Supplementary Material online). Finally, site 370 is within SRS5. Some of these sites also had amino acid changing properties that could affect the polarity, hydropathy, isoelectric point, chemical composition, volume, secondary structure, and refractivity of the CYP2H sequences (supplementary table S9, Supplementary Material online).

CYP2J: For "CYP2J_1," model M2a indicated approximately 7% of sites under positive selection ($\omega 2 = 3.591$) whereas model M8 indicated 9% ($\omega = 3.162$) (supplementary table S10, Supplementary Material online). Both models identified positively selected sites located within SRS0 (22, 43, and 46), SRS1 and HEM (73, 75 and 76) and SRS2_SRS3 (182, 185, 186, 199, 204, 207, and 208)

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

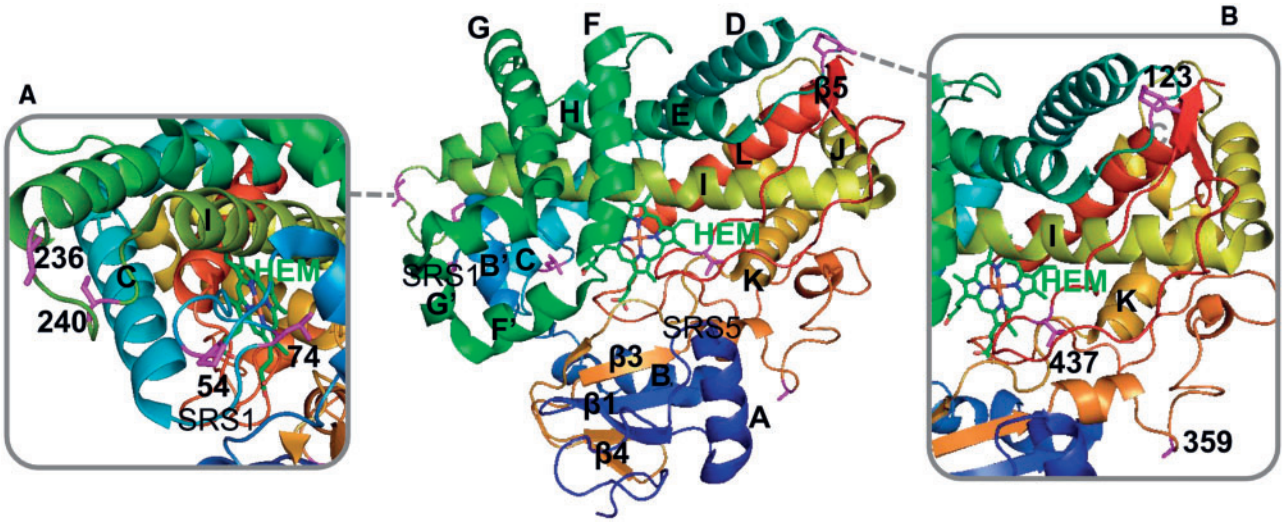


Fig. 6.—3D analyses of sites detected to be under positive selection in the avian CYP2D. The CYP2D 3D structure from PDB (3TDA human) was used to map the positively selected sites found. The sites are shown as pink sticks. The regions corresponding to the helices are named with the corresponding capital letter and the β-structures are named with a β followed by a number. HEM represents the heme group. One SRS region is represented, the SRS1 (on the left).

(supplementary table S11, Supplementary Material online). These sites accounted for approximately 76% of the positively selected sites that were found. Changes at some of these sites likely changed properties such as chemical composition (43), hydropathy (182), and volume (185) (supplementary table S11, Supplementary Material online). For "CYP2J_2," model M2a indicated approximately 2% of sites under positive selection ($\omega 2 = 3.952$) whereas model M8 indicated 5% ($\omega = 2.346$) (supplementary table S12, Supplementary Material online). Both models identified only two sites (58 and 59) under positive selection located within the SRS0. For site 59, the PRIME analysis suggests that the mutation would cause a shift in polarity. For "CYP2J_3," only model M8 indicated approximately 10% of sites under positive selection ($\omega = 1.360$—supplementary table S13, Supplementary Material online). Only one site (21) was positively selected and no amino acid property was selected. For "CYP2J_4" model M2a indicated approximately 2% of sites under positive selection ($\omega 2 = 3.033$) whereas model M8 indicated 5% ($\omega = 2.038$) (supplementary table S14, Supplementary Material online). Two (201 and 250) of the five positively selected sites, identified by both models, were located within SRS2_SRS3 (supplementary table S15, Supplementary Material online). For this last site, refractivity was the amino acid changing property (supplementary table S15, Supplementary Material online). Finally, for "CYP2J_5" only model M8 indicated approximately 1% of sites under positive selection ($\omega = 2.124$— supplementary table S16, Supplementary Material online). Two sites (65 and 196) were selected. However, only site 196 was located in the SRS2_SRS3, and selective changes would have affected its isoelectric point.

CYP2K: For "CYP2K_1," both models indicated approximately 3% of sites under positive selection (M2a: $\omega 2 = 3.679$ and M8: $\omega = 3.088$—supplementary table S17, Supplementary Material online) and nine sites were identified (supplementary table S18, Supplementary Material online). From these, four sites were within important regions of the enzyme: 104 (SRS1 and HEM) and 217, 230, and 240 (SRS2_SRS3). For these sites, the following changing properties were identified: refractivity (104), volume (217), and isoelectric point (230 and 240) (supplementary table S18, Supplementary Material online). For "CYP2K_2," model M2a indicated approximately 1% of the sites under positive selection ($\omega 2 = 3.341$) whereas model M8 indicated 3% ($\omega = 1.515$) (supplementary table S19, Supplementary Material online), especially sites 277 and 321.

CYP2AC: Model M2a indicated approximately 12% of sites under positive selection ($\omega 2 = 1.628$) whereas model M8 indicated 17% ($\omega = 1.517$) (supplementary table S20, Supplementary Material online). Only one site (92) was identified with PP > 95% (model M8). This site was located in the SRS1 and HEM regions and the property volume was pointed as acting in this site.

The remaining subfamilies CYP2R (supplementary table S21, Supplementary Material online), CYP2U (supplementary table S22, Supplementary Material online), CYP2W (data sets "CYP2W_1"—supplementary table S23, Supplementary Material online and "CYP2W_2"—supplementary table S24, Supplementary Material online), CYP2AF (supplementary table S25, Supplementary Material online), and CYP2J (only data set named as "CYP2J"—supplementary table S26,

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

Supplementary Material online) showed no significant evidence of positive selection.

Selection of CYP2 Has Been Branch-Specific in Species with Shared Traits

The branch-specific analyses, applied to CYP2 subfamilies with evidence of positive selection (from site-models), revealed that "CYP2H" (ωcarnivorous = 0.570, ωomnivorous = 0.490, and ωherbivorous = 0.555) and "CYP2J_2" (ωcarnivorous = 0.265, ωomnivorous = 0.356, and ωherbivorous = 0.393) are evolving differently in birds with distinct carnivorous, omnivorous, and herbivorous feeding habits (fig. 7). The strength of selection is also variable among birds occupying distinct habitats (moist, semi-moist and/or dry) in six of the CYP2 subfamily data sets ("CYP2D," "CYP2H," "CYP2J_2," "CYP2J_3," "CYP2J_5," and "CYP2K_1") (fig. 7). "CYP2D," "CYP2H," and "CYP2J_3" are evolving differently in migratory and non-migratory birds (supplementary figs. S1–S11, Supplementary Material online). Only the "CYP2AC"(ω = 0.353), "CYP2C" (ω = 0.362), "CYP2J_1" (ω = 0.450), "CYP2J_4" (ω = 0.252), and "CYP2K_2" (ω = 0.178) subfamily data sets are evolving at the same rate in all bird groups regardless of their distinct feeding habits, habitats and migratory behaviors (fig. 7).

CYP2 Gene Subfamily Numbers Vary According to Migration and Feeding Habits

The LDA constructed a single discriminant function (fig. 8). The success of classification, estimated by cross-validation, was low (58% global, 67% migratory, and 48% nonmigratory). This was due to the large degree of similarity among the number of CYP2 genes present in the two classes (migratory and non-migratory), as is apparent from the average variable scores per group (fig. 8A). The minor differences found between the two classes were due to CYP2 genes scoring in the extremes of the discriminant function, with migratory bird scores tending slightly towards negative values whereas non-migratory are closer to the positive end of the function (fig. 8B). However, a Kolmogorov–Smirnoff test applied to the frequencies of scores from the linear discriminant function (fig. 8B) showed a significant difference between the distribution of each class (P value < 0.01). Supplementary table S27, Supplementary Material online shows that among all these genes, CYP2D, CYP2U, CYP2H, and CYP2AF are the main ones responsible for negative scores (migratory) whereas CYP2K is responsible for positive scores (non-migratory).

Furthermore, differences in the number of CYP2 genes were also observed when analysing feeding habits in association with the migratory behavior of birds (fig. 9). Migratory carnivores and herbivores have less CYP2 genes than migratory omnivores, but these differences were only significant between migratory carnivores and migratory omnivores (Mann–Whitney test, P value < 0.05). Non-migratory carnivores have less CYP2 genes than non-migratory omnivores and herbivores, but none of these differences resulted to be significant. Globally, omnivores have a higher number of CYP2 genes than specialist birds (carnivorous and herbivorous birds), except for non-migratory herbivores, which have a

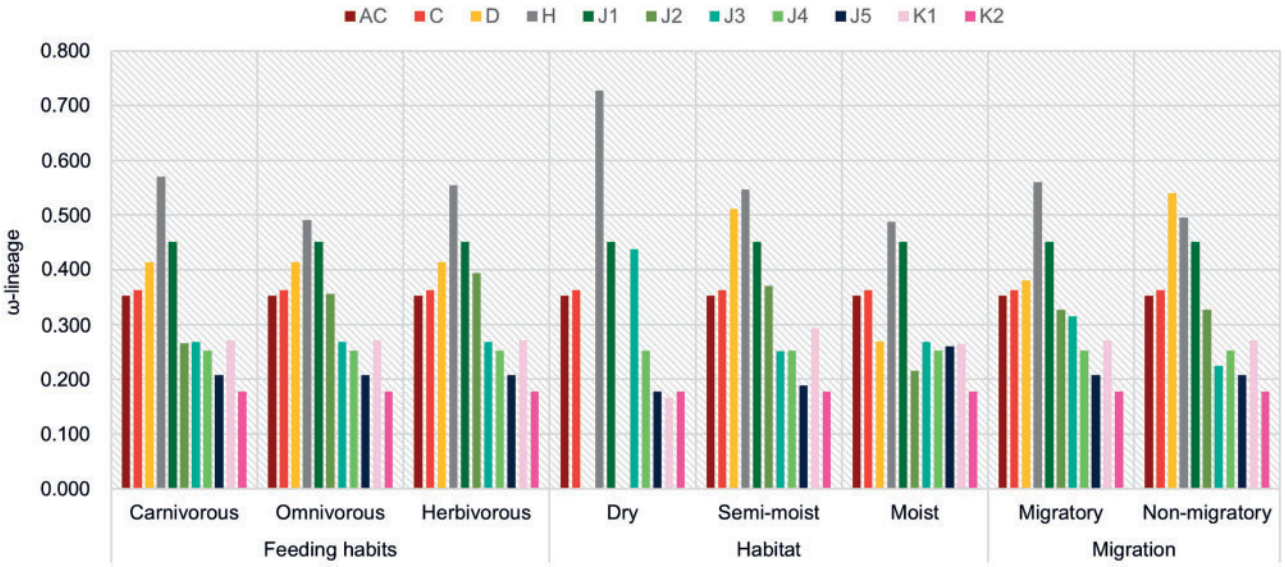


Fig. 7.—ω-ratio variations according to distinct avian feeding habits, habitats and migratory behaviors. The ω values represented arise from the hypothesis that best fits each CYP2 subfamily data set according to the branch-specific LRT analyses (P value < 0.05) (see supplementary figs. S1–S11, Supplementary Material online for more methodological details). CYP2 subfamily code abbreviations indicate each one of the data sets used: AC–"CYP2AC," C–"CYP2C," D–"CYP2D," H–"CYP2H," J1–"CYP2J_1," J2–"CYP2J_2," J3–"CYP2J_3," J4–"CYP2J_4," J5–"CYP2J_5," K1–"CYP2K_1," and K2–"CYP2K_2."

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

number of CYP2 genes similar to non-migratory omnivores (fig. 9). The only significant differences were detected between migratory omnivores and both types of carnivores (Mann–Whitney test, P value < 0.05) (fig. 9).

Discussion

The avian CYP2 subfamilies corresponded well with previous classifications (Watanabe et al. 2013), including the finding that all avian species have only a single gene member in the CYP2D, CYP2R, and CYP2U subfamilies (fig. 1). CYP2J, CYP2W, and CYP2K were the largest subfamilies in the 48 avian genomes analyzed, with two to seven duplicated genes (fig. 1). Although it had previously been suggested that the *CYP2W1* and *CYP2W2* genes had duplicated only in the Galloanserae lineage (Watanabe et al. 2013), our results covering all the three main avian groups (Palaeognathae, Galloanserae, and Neoaves) clarified that there have been several CYP2W duplication events (fig. 1). Our identification of avian CYP2F, CYP2G, and CYP2AF subfamilies indicates that the CYP2F and CYP2G subfamilies are not mammalian-specific as was previously suggested (Kirischian et al. 2011) and that the CYP2AF has not been lost in the avian lineage, contrarily to what has been previously hypothesized (Watanabe et al. 2013).

Some of our most striking results are the presence or absence of one or many CYP families in closely related species. The Adelie penguin and the closely related emperor penguin have similar habitats (both live exclusively in the Antarctic region), feeding habits (carnivores), and lifestyles (colonial, social, specialized for swimming) (Williams 1995) but the number of CYP2 subfamilies between them was striking (nine and only three, respectively—fig. 3). This is an intriguing scenario because it was expected that these species facing similar selective pressures would present a similar number of CYP2 subfamilies. However, according to our approach the emperor penguin appears to have lost many subfamilies that the Adelie penguin still has and shares with other more distantly related bird species. In order to exclude possible false-negatives due to the strict criteria of our identification approach, we performed exploratory tBLASTn searches against the emperor penguin genome with less stringent criteria and using the Adelie penguin CYP2 sequences as query. We retrieved some partial sequences (ranging between 177 and 849 bp) possibly representing the CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, CYP2W, and CYP2U subfamilies. However,

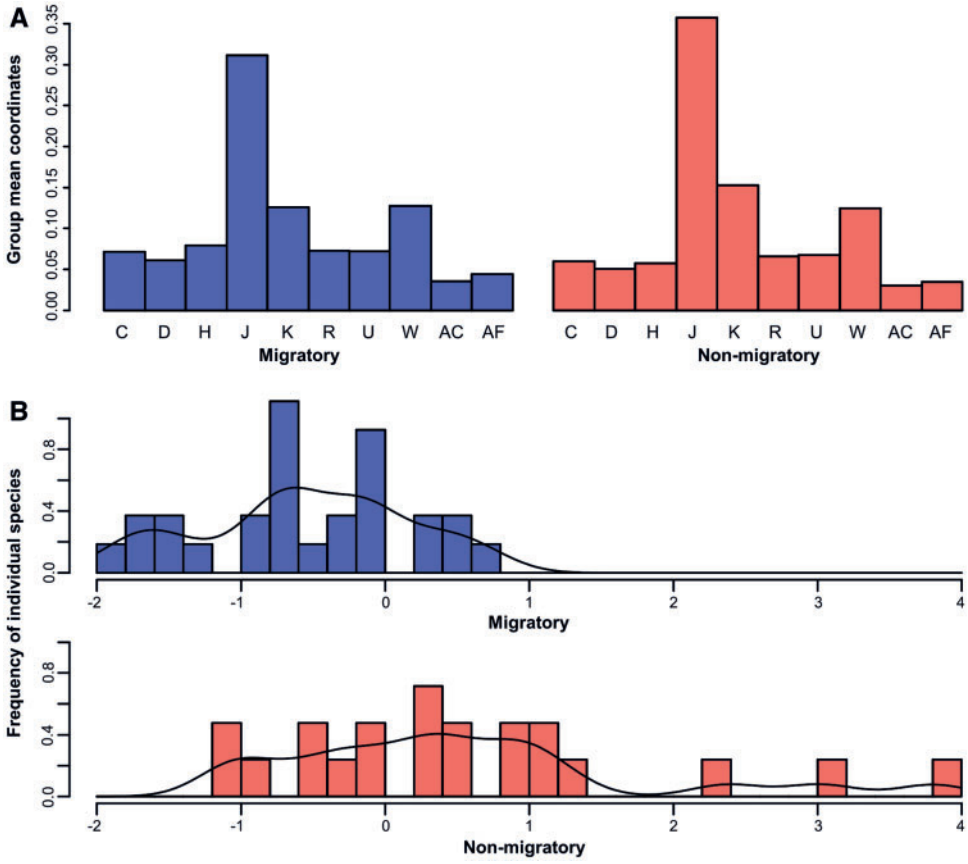


Fig. 8.—Results of the linear discriminant function. (*A*) Average scores per class. (*B*) Individual species scores per class. The black line represents a nonparametric fit of frequency distributions.

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

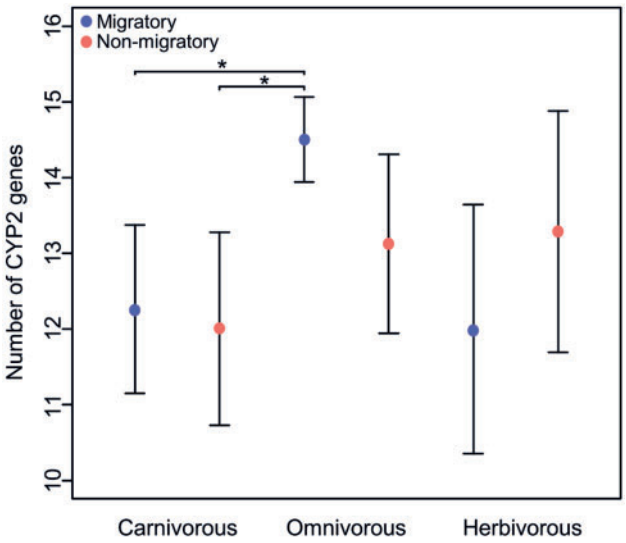


Fig. 9.—Comparison of the number of CYP2 genes in birds with different migratory behaviors and feeding habits. Ploted are the mean ± 95% confidence interval. Asterisks indicate significant differences in pairwise comparisons performed by Mann–Whitney tests. Species trait classifications are in supplementary table S2, Supplementary Material online.

due to their short lengths (as opposed to complete CYP2 genes that encompass approximately 1,500 bp), the accurate assignment to a CYP subfamily becomes compromised. Thus, such short sequences were not considered for further detailed evolutionary analyses. Conversely, CYP2F is only present in grey-crowned crane and CYP2G in chimney swift. This would be consistent with other studies. For example, a study of 200 humans found that a functional CYP2G allele was also uncommon in humans (detected in only 11.6% of the individuals) (Sheng et al. 2000). We cannot exclude the possibility that these genes are present in the gaps of avian genome assemblies, or even that evidence of their presence in other birds can be missed, as demonstrated above, by our threshold of requiring at least 1,125 bp when searching for avian CYP2 nucleotide sequences.

Our selection analyses results are consistent with some findings in avian and nonavian species. The residue at position 369 of the avian "CYP2C" data set was within SRS5 and corresponds to site 364 in *Oryctolagus cuniculus* (rabbit) *CYP2C3v*, where the mutation T364S has been linked with changes in progesterone region selectivity (Richardson and Johnson 1994). The chicken CYP2C gene is activated by the Chicken Xenobiotic Receptor, supporting its role in xenobiotic metabolism (Baader et al. 2002). Thus, it is plausible that the positively selected sites (239, 281, 369, and 453) leading to changes in amino acid properties might have provided an important adaptation by facilitating the efficient inactivation and removal of several xenobiotic compounds in birds.

CYP2D is present in several mammalian species (e.g., rodents, primates, rabbit, and horse), and has been linked with feeding habits and with metabolizing plant toxins such as alkaloids (Yasukochi and Satta 2011). Therefore, the positively selected sites 54 and 74 found in the avian CYP2D subfamily, located within the SRS1 and HEM functional regions, could be particularly advantageous for an efficient dietary detoxification (Yasukochi and Satta 2015). Our lineage-specific analyses of CYP2D suggest similar impacts among birds with distinct feeding habits. The effect of changing amino acid properties for some of the positively selected sites (123, 236, 359) located outside of the active site areas of the avian CYP2D subfamily could possibly be related with global protein folding or substrate recognition (da Fonseca et al. 2007).

CYP2H enzymes are involved in reactions of epoxygenation (Kanetoshi et al. 1992) and they are the major phenobarbital-inducible enzymes in the chicken liver (Hu 2013). Agonist for the CYP2H are the drugs dexamethasone and metyrapone and also the compounds okadaic acid, pregnenolone16 alpha-carbonitrile, and squalestatin1 (Ourlin et al. 2000). The substrate of these enzymes is arachidonic acid (Kanetoshi et al. 1992), which implies that positively selected sites located within the SRS1 and HEM regions of the avian CYP2H subfamily (site 102), within the SRS4 and HEM binding site (305), and in the HEM binding region and in the SRS5 (site 365) might have an impact on the activity of this enzyme in birds. Since CYP2H has evolved differently among different avian groups, it is most likely that the changes have distinct adaptive relevance possibly involving feeding habits, habitats, and migratory behaviors.

The greater number of gene duplications and large variability in the amino acid patterns among the different copies of CYP2J genes among birds could also be related to different habitats and feeding adaptations. Similar duplications events have been detected in bactrian camels, which are hypothesized to be linked with the importance of CYP2J in the conversion of arachidonic acid into 19(S)-HETE—a potential vasodilator of renal preglomerular vessels—that stimulates water reabsorption (Jirimutu et al. 2012). CYP2J also has epoxygenase activity and can convert arachidonic acid into epoxyeicosatrienoic acids (EETS) that have antihypertensive vasodilatory properties (Yu et al. 2000). Thus, it has been hypothesized that an increased number of CYP2J copies would increase water absorption and could influence survival in dry conditions (Jirimutu et al. 2012). The large number of avian CYP2J copies and the extraordinary high degree of positive selection detected in the SRS suggest that this subfamily might be important to adaptation to distinct habitats by using water more efficiently. The genes from the "CYP2J_2," "CYP2J_3," and "CYP2J_5" subfamily data sets could have a particular role in this process as we found: 1) distinct ω for "CYP2J_2" among carnivorous, omnivorous, herbivorous, land (semi-moist), and water (moist) birds; 2) distinct ω for "CYP2J_3" among land (dry and semi-moist), water (moist), migratory,

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

and non-migratory birds, and 3) distinct ω for "CYP2J_5" among land (dry and semi-moist) and water (moist) birds.

Although we have identified sites under positive selection in the avian CYP2K and CYP2AC subfamilies, further studies are necessary to infer the impact of these substitutions, since their function remains little known.

The absence of positive selection in CYP2R and CYP2U subfamilies could be explained by their essential role in the metabolism of vitamin D and arachidonic acid. These genes have conserved synteny between birds and humans (Watanabe et al. 2013). Similarly, the remaining subfamilies without positive selection (CYP2W and CYP2AF) also can have essential functions in the metabolism of endogenous compounds, being strongly adapted to their substrates.

For the LDA analyses of CYP2 genes, despite the low success of classification of genes, this analysis provides an opportunity for the future, when a greater number of avian genomes will become available, because it suggests that some CYP2 subfamilies, including CYP2D, CYP2U, CYP2H, and CYP2AF, are more related to migratory species. Our analysis also revealed that the variation in the number of CYP2 genes is related to different feeding habits and migratory behaviors of birds. This suggests that the higher number of CYP2 genes found in avian migratory omnivores might be related with their need to adapt to a wider variety of environments and food resources, and thus the higher exposure to several toxins, would require a more efficient detoxification capacity (Rainio et al. 2012).

While our study provides significant advances to our understanding of avian CYP2 evolution, the currently available genome scaffold lengths limit the number of CYP2 sequences that can be confidently classified. Ongoing efforts to increase the scaffold lengths of bird genomes will enhance our understanding of avian CYP2 identification and adaptation.

Conclusions

To our knowledge, this is the first study of avian CYP2 subfamilies that includes representatives of the three avian evolutionary groups: Palaeognathae, Galloanserae, and Neoaves (Zhang et al. 2014a). We identified 12 CYP2 subfamilies in 48 avian genomes and showed that some of the CYP2 genes that were previously described as being lineage-specific, such as CYP2K and CYP2W, are present in representatives of all the avian groups. Additionally, we demonstrated the presence of the CYP2F and CYP2G subfamilies in some avian genomes. From our comparative analyses we updated our knowledge of the SRS of CYPs, and identified several new regions (SRS0, SRS2_SRS3, and SRS3.1). We identified several significant signatures of positive selection in the six avian CYP2 subfamilies (CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, and CYP2AC), some of which are located in relevant SRS- and heme-binding areas (HEM) that influence CYP2 structure and function. The six CYP2 subfamilies that showed positive selection had sites under positive selection in HEM and in one or both SRS1 and SRS3. Of the six, only the CYP2C and CYP2H subfamilies had sites under positive selection in SRS5, suggesting that these two subfamilies may be under similar evolutionary pressures in this enzyme region, that allow them to phenotypically adapt and acquire similar substrate affinities. The positive selected sites in these avian CYP2 subfamilies likely have helped them adapt to distinct chemical compounds in new habitats with distinct food resources, and facilitate the dispersion and evolutionary success of birds.

Supplementary Material

Supplementary figures S1–S12 and supplementary tables S1–S27 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We owe special thanks to Tibisay Escalona, Aldo Barreiro, Rui Borges, and Rosalina Goulão for their helpful discussions and support. The authors are thankful to the Associate Editor Dr José Pereira-Leal and the anonymous reviewers for their valuable comments and suggestions. D.A. and I.K. were funded with a PhD grant from Fundação para a Ciência e a Tecnologia (FCT) (SFRH/BD/79766/2011 and SFRH/BD/48518/2008, respectively). E.D.J. was supported by the Howard Hughes Medical Institute (HHMI). S.J.O. was supported as Principal Investigator by Russian Ministry of Science Mega-grant no.11.G34.31.0068. A.A. was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and European Regional Development Fund (ERDF) in the framework of the programme PT2020, and the FCT project PTDC/AAG-GLO/6887/2014.

Literature Cited

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 18:1585–1592.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950–958.

Antunes A, Ramos MJ. 2007. Gathering computational genomics and proteomics to unravel adaptive evolution. Evol Bioinform. 3:207–209.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195–201.

Atchley WR, Zhao J, Fernandes AD, Druke T. 2005. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A. 102:6395–6400.

Baader M, Gnerre C, Stegeman JJ, Meyer UA. 2002. Transcriptional activation of cytochrome P450 *CYP2C45* by drugs is mediated by the chicken xenobiotic receptor (CXR) interacting with a phenobarbital response enhancer unit. J Biol Chem. 277:15647–15653.

Biasini M, et al. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 42:W252–W258.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7:88.

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
Conant GC, Wagner GP, Stadler PF. 2007. Modeling amino acid substitution patterns in orthologous and paralogous genes. Mol Phylogenet Evol. 42:298–307.
Csuros M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912.
da Fonseca RR, Antunes A, Melo A, Ramos MJ. 2007. Structural divergence and adaptive evolution in mammalian cytochromes P450 2C. Gene 387:58–66.
Dalloul RA, et al. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biol. 8(9):e1000475.
Danielson PB. 2002. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr Drug Metab. 3:561–597.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9:772.
DeLano WL. 2002. The PyMOL Molecular Graphics System. Version 1.5.0.4: Schrödinger, LLC, Cambridge. Available from: www.pymol.org
Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Farris JS. 1977. Phylogenetic analysis under Dollo's Law. Syst Biol. 26:77–88.
Flicek P, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42:D749–D755.
Gotoh O. 1992. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. J Biol Chem. 267:83–90.
Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 27:221–224.
Guengerich FP. 2008. Cytochrome p450 and chemical toxicology. Chem Res Toxicol. 21:70–83.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.
Hu SX. 2013. Effect of age on hepatic cytochrome P450 of Ross 708 broiler chickens. Poult Sci. 92:1283–1292.
International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.
Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.
Jirimutu, et al. 2012. Genome sequences of wild and domestic bactrian camels. Nat Commun. 3:15647. doi: 10.1038/ncomms2192.
Kanetoshi A, Ward AM, May BK, Rifkind AB. 1992. Immunochemical identity of the 2,3,7,8-tetrachlorodibenzo-p-dioxin- and beta-naphthoflavone-induced cytochrome P-450 arachidonic acid epoxygenases in chick embryo liver: distinction from the omega-hydroxylase and the phenobarbital-induced epoxygenase. Mol Pharmacol. 42:1020–1026.
Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.
Kirischian N, McArthur AG, Jesuthasan C, Krattenmacher B, Wilson JY. 2011. Phylogenetic and functional analysis of the vertebrate cytochrome p450 2 family. J Mol Evol. 72:56–71.
Konstandi M, Johnson EO, Lang MA. 2014. Consequences of psychophysiological stress on cytochrome P450-catalyzed drug metabolism. Neurosci Biobehav Rev. 45C:149–167.
Maldonado E, Sunagar K, Almeida D, Vasconcelos V, Antunes A. 2014. IMPACT_S: integrated multiprogram platform to analyze and combine tests of selection. PLoS One 9:e96243.
Martin DP, et al. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462–2463.
McClellan DA, et al. 2005. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. Mol Biol Evol. 22:437–455.
Nebert DW. 2000. Suggestions for the nomenclature of human alleles: relevance to ecogenetics, pharmacogenetics and molecular epidemiology. Pharmacogenetics 10:279–290.
Nebert DW, Gonzalez FJ. 1987. P450 genes: structure, evolution, and regulation. Annu Rev Biochem. 56:945–993.
Nebert DW, Russell DW. 2002. Clinical importance of the cytochromes P450. Lancet 360:1155–1162.
Nebert DW, Wikvall K, Miller WL. 2013. Human cytochromes P450 in health and disease. Philos Trans R Soc Lond B Biol Sci. 368:20120431.
Nelson DR. 1998. Cytochrome P450 nomenclature. Methods Mol Biol. 107:15–24.
Nelson DR. 2003. Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. Arch Biochem Biophys. 409:18–24.
Nelson DR. 2009. The cytochrome p450 homepage. Hum Genomics. 4:59–65.
Nelson DR, et al. 1993. The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. DNA Cell Biol. 12:1–51.
Nelson DR, et al. 2004. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14:1–18.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
Ourlin JC, Baader M, Fraser D, Halpert JR, Meyer UA. 2000. Cloning and functional expression of a first inducible avian cytochrome P450 of the CYP3A subfamily (CYP3A37). Arch Biochem Biophys. 373:375–384.
Palmer G, Reedijk J. 1991. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature of electron-transfer proteins. Recommendations 1989. Biochim Biophys Acta. 1060:599–611.
Pennell MW, et al. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216–2218.
Pochapsky TC, Kazanis S, Dang M. 2010. Conformational plasticity and structure/function relationships in cytochromes P450. Antioxid Redox Signal. 13:1273–1296.
Pond SL, Frost SD. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533.
Poulos TL, Finzel BC, Howard AJ. 1987. High-resolution crystal structure of cytochrome P450cam. J Mol Biol. 195:687–700.
R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.
Rainio MJ, Kanerva M, Wahlberg N, Nikinmaa M, Eeva T. 2012. Variation of basal EROD activities in ten passerine bird species–relationships with diet and migration status. PLoS One 7:e33926.
Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 3:217–223.
Richardson TH, Johnson EF. 1994. Alterations of the regiospecificity of progesterone metabolism by the mutagenesis of two key amino

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024

acid residues in rabbit cytochrome P450 2C3v. J Biol Chem. 269:23937–23943.
Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19:101–109.
Saxena A, Tripathi KP, Roy S, Khan F, Sharma A. 2008. Pharmacovigilance: effects of herbal components on human drugs interactions involving cytochrome P450. Bioinformation 3:198–204.
Scott JG, Wen Z. 2001. Cytochromes P450 of insects: the tip of the iceberg. Pest Manag Sci. 57:958–967.
Sheng J, Guo J, Hua Z, Caggana M, Ding X. 2000. Characterization of human CYP2G genes: widespread loss-of-function mutations and genetic polymorphism. Pharmacogenetics 10:667–678.
Sullivan RJ, Hagen EH, Hammerstein P. 2008. Revealing the paradox of drug reward in human evolution. Proc Biol Sci. 275:1231–1241.
Suzuki Y, Nei M. 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol. 18:2179–2185.
Suzuki Y, Nei M. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol Biol Evol. 19:1865–1869.
Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol. 20:18–20.
Thomas JH. 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3:e67.
Warren WC, et al. 2010. The genome of a songbird. Nature 464:757–762.
Watanabe KP, et al. 2013. Avian cytochrome P450 (CYP) 1-3 family genes: isoforms, evolutionary relationships, and mRNA expression in chicken liver. PLoS One 8:e75689.
Williams TD. 1995. The Penguins: Spheniscidae: Bird families of the World. New York: Oxford University Press p.295. ISBN: 019854667X.
Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041–1051.
Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered. 92:371–373.
Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol Phylogenet Evol. 26:1–7.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol. 51:423–432.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.
Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 25:568–579.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.
Yasukochi Y, Satta Y. 2011. Evolution of the CYP2D gene cluster in humans and four non-human primates. Genes Genet Syst. 86:109–116.
Yasukochi Y, Satta Y. 2015. Molecular evolution of the CYP2D subfamily in primates: purifying selection on substrate recognition sites without the frequent or long-tract gene conversion. Genome Biol Evol. 7:1053–1067.
Yu Z, et al. 2000. Increased CYP2J expression and epoxyeicosatrienoic acid formation in spontaneously hypertensive rat kidney. Mol Pharmacol. 57:1011–1020.
Zhang G, Li B, Li C, Gilbert MTP, Jarvis E. 2014. The Avian Genome Consortium, Wang J. 2014. The avian phylogenomic project data. In: GigaScience Database.
Zhang G, Li C, Li Q, Li B, Larkin DM, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346:1311–1320.

Associate editor: José Pereira-Leal

Downloaded from https://academic.oup.com/gbe/article/8/4/1115/2574007 by guest on 17 April 2024