Genome analysis

Phylogenetic analysis of three multi-gene families across primates: CYPs, GSTs, and CESs

Jade Westlake [1,*], and Dr. Lars Jermin [2,*]

[1] School of Mathematical and Statistical Sciences, University of Galway, Galway, H91 TK33, Ireland
[2] School of Mathematical and Statistical Sciences, University of Galway, Galway, H91 TK33, Ireland.

*j.westlake1@universityofgalway.ie

Abstract

The Glutathione S-Transferase (GST), Cytochrome P450 (CYP), and Carboxylesterase (CES) gene families encode key enzymes involved in the metabolism and detoxification of endogenous and exogenous compounds. Despite their clinical relevance and well-characterized roles in drug metabolism and drug-resistance, their evolutionary trajectories across primates remain poorly understood. This study investigates the diversification and evolution of GST, CYP, and CES gene families across 13 primate species. 575 GST, 272 CES, and 1,285 CYP genes were identified through BLAST-based homology searches. Multiple sequence alignments were performed using MAFFT and viewed on Jalview and Aliview, and phylogenetic trees were constructed with IQ-TREE2 and FastTree to generate high-resolution phylogenies. Comparative genomic analysis revealed substantial interspecies variation in gene copy number and lineage-specific expansions, highlighting genetic diversity in these families even among closely related primates. Variation encompassed differences in gene copy number, isoform diversity, and retention of canonical gene forms, with humans showing notable expansion in CES4 isoforms. This genetic variation in GST, CES, and CYP gene families is known to contribute to differences in drug response.

1 Introduction

Glutathione S-transferase (GST), cytochrome P450 (CYP), and carboxylesterases (CES) are three multi-gene families found in a wide range of species, from bacteria and fungi to plants and animals [22]. Enzymes belonging to these three multi-gene families have similar functionality in drug metabolism, detoxification of xenobiotics, and prodrug activation/deactivation. Organisms have been exposed to a wide array of foreign chemical compounds since ancient times. These foreign substances are collectively termed xenobiotics which can be toxic and include either ancient substances i.e. plant and fungal metabolites, such as phenolic compounds and aflatoxins, as well as reactive oxygen species like superoxide radicals and hydrogen peroxide, or more recent substances, i.e. foreign substances generated from human activities over the last two centuries have introduced many synthetic chemicals. The capacity to metabolize and detoxify these harmful agents, whether produced internally or encountered in the environment, represents a critical evolutionary adaptation to survival [28].

There are four stages in xenobiotic metabolism: influx through membrane transporter proteins, biotransformation through Phase I reactions, conjugation through Phase II metabolism, and elimination via Phase III transport systems [5]. Phase I metabolism, primarily catalyzed by CYPs, introduces or exposes functional groups such as hydroxyl, thiol, amino, or carboxyl groups to increase the compound's polarity and water solubility. These modifications create reactive sites for subsequent Phase II reactions. CES are phase I enzymes that catalyze the hydrolysis of endogenous and exogenous compounds, including esters, thioesters, carbamates, and amides. Mammalian CES enzymes play key roles in metabolizing a wide range of substrates such as environmental toxins, therapeutic drugs, and prodrugs, thereby influencing both detoxification and drug activation pathways [3]. CYPs are a major Phase I superfamily of membrane-bound hemoproteins widely expressed across almost all tissues, with particularly high concentrations in the liver, small intestine, and kidneys [6]. CYP enzymes are essential for metabolizing a wide range of lipophilic substances, including drugs, environmental chemicals, and dietary compounds, by catalyzing their oxidative biotransformation to facilitate detoxification. CYPs also play vital roles in endogenous pathways such as steroid hormone regulation, bile acid synthesis, vitamin metabolism, and cholesterol biosynthesis [5].

Phase II involves conjugation of the xenobiotic, or Phase I metabolite, with hydrophilic molecules such as glutathione or glucuronic acid, further enhancing solubility to facilitate excretion [28]. GSTs are a major family of phase II detoxification enzymes primarily located in the cytosol, where they catalyze the conjugation of electrophilic compounds to glutathione (GSH), enhancing their solubility. GSTs also exhibit diverse functions, including peroxidase and isomerase activities, modulation of signaling pathways, such as inhibition of Jun N-terminal kinase to protect cells from oxidative stress, and non-catalytic binding to a variety of endogenous and exogenous ligands, contributing to both detoxification and bioactivation of xenobiotics [28, 26]. Finally, in Phase III, these water-soluble conjugates are actively transported out of the cell via specialized efflux transporters [5].

CYP, CES, and GST enzymes are believed to have originated over 2-3 billion years ago, likely within early prokaryotes such as bacteria and archaea, where they have since continuously evolved and diversified; however, CES has been described to have emerged later than GSTs and CYPs [9]. As metazoans emerged, these enzymes adapted to perform vital endogenous functions, including fatty acid metabolism. During the Devonian period, approximately 400 million years ago, the transition of life to terrestrial environments introduced new selective pressures. Animals and insects encountered plant-produced toxins evolved to deter herbivory, which in turn drove the evolution of specialized enzymes for xenobiotic detoxification, allowing organisms to neutralize these harmful compounds effectively. Another significant expansion of these genes is believed to have occurred following the Permian-Triassic mass extinction around 250 million years ago, as new ecological niches and selective pressures emerged from this global event [15].

Dietary and ecological pressures have shaped the evolution and diversification of GSTs, CYPs, and CESs. Our study is a phylogenetic analysis of CYP, GST, and CES evolution across primate species with diverse ecological niches and dietary habits (Table 1), ranging from frugivorous gibbons in Southeast Asian rainforests to omnivorous humans distributed globally. This ecological diversity is central to the selective pressures that have shaped the evolution of these detoxification enzyme families. Species consuming diets rich in plant secondary metabolites, such as fruits, leaves, and seeds, face a wider variety of xenobiotic compounds. For example, frugivorous species like *Pongo pygmaeus* and *Nomascus leucogenys* primarily consume fruits and leaves that contain numerous potentially toxic phytochemicals, necessitating an adequate detoxification system. Similarly, omnivorous species such as *Homo sapiens*, *Pan troglodytes*, and *Macaca mulatta* incorporate a broad dietary spectrum including animal prey, insects, and human-associated foods, further diversifying their xenobiotic exposure. Reflecting this ecological and dietary variability, our results demonstrate significant genetic variation in gene counts and diversification of GSTs, CYPs, and CESs across the primate species examined.

These three gene families have been the focus of numerous clinical studies due to their variable expression, which leads to inconsistent drug responses among patients. Patients can be placed in separate groups based on their inherited rate of drug metabolism, which can significantly impact a patient's response to drug therapy. Therefore, understanding the mechanisms underlying a drug's action on these enzymes is essential for choosing the most efficient therapy [6]. All three families have been studied as drug targets; however, CESs have been largely overlooked at the clinical level [3], despite accounting for approximately 1 percent of the entire liver proteome, and contributing to 80–95 percent of the liver's total hydrolytic activity. CES enzymes play a crucial role in the metabolism of a wide range of substances, including drugs (particularly ester prodrugs), pesticides, environmental pollutants, and endogenous compounds. Carboxylesterase-mediated hydrolysis also plays an important role in the disposition of several widely prescribed therapeutic agents from a diverse range of drug classes including: antiplatelet drugs, angiotensin converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), central nervous system stimulants (CNS stimulants), antiviral agents, and immunosuppressants [8] for example; recent studies have shown that CES1 functions as a cholesteryl ester hydrolase involved in lipid metabolism in human macrophages and hepatocytes, highlighting its potential as a drug target for treating metabolic diseases such as diabetes and atherosclerosis [8].

CYPs have been the primary focus of research among these gene families [5]. Over the last 30 years, it has become clear that genetic variability of CYPs in the patient is highly relevant in terms of drug response. CYPs exhibit genetic polymorphisms with multiple allelic variants, demonstrating frequencies varying between different populations and ethnicities [5]. Of the total 57 isozymes discovered to date, 6 of these are responsible for 90 percent of drug metabolism, including CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4 [6]. For several diseases, it is already a strategy to reduce the concentrations of CYP products to cure the disease. Several of these involve interfering with steroid metabolism, which is dominated by CYP. An example is targeting CYP5A1 (Thromboxane Synthase) in the treatment of several cardiovascular diseases with the aim to lower thromboxane levels and inhibit platelet aggregation [7].

Table 1. Ecological details (habitat and diet) of 13 primate species used in this study.

Species	**Habitat and Diet**
Homo sapiens	Global distribution. Omnivorous diet including meat, grains, fruits, vegetables, and dairy.
Pan troglodytes	Central and West African rainforests and savannas. Omnivorous – mostly fruit, leaves, seeds, insects, occasional meat.
Pan paniscus	Congo Basin rainforests. Frugivorous – mostly fruit, with leaves, flowers, and small animals.
Gorilla beringei beringei	East African montane forests. Herbivorous – leaves, stems, shoots, and some fruit.
Gorilla gorilla	Central African lowland forests. Herbivorous – primarily fruit, leaves, and seeds.
Pongo pygmaeus	Bornean tropical and swamp forests. Frugivorous – mainly figs, with bark and insects.
Pongo abelii	Northern Sumatran rainforests. Frugivorous – fruit, leaves, flowers, and insects.
Hylobates moloch	Rainforests of Java. Frugivorous – fruits, leaves, and insects.
Nomascus leucogenys	Forests of Laos, Vietnam, and southern China. Frugivorous – fruits, leaves, and insects.
Symphalangus syndactylus	Sumatra and Malay Peninsula rainforests. Omnivorous – fruits, leaves, flowers, and insects.
Callithrix jacchus	Northeastern Brazil forests. Omnivorous – tree sap, insects, fruit, and small animals.
Macaca mulatta	Forests and urban areas across South and Southeast Asia. Omnivorous – fruits, seeds, insects, and human food.
Microcebus murinus	Madagascar dry forests. Omnivorous – fruits, insects, flowers, and small vertebrates.

Chemo-resistance remains a major challenge in cancer therapy that can arise through various mechanisms, including drug inactivation, impaired apoptosis, deregulation of the cell cycle and checkpoints, as well as acquired genetic mutations and epigenetic modifications. Among these, the inactivation of chemotherapeutic agents by detoxification enzymes, particularly by GSTs, has been widely investigated. Many cancer types exhibit distinct GST expression profiles that offer opportunities for targeted therapies. The catalytic activity of GST enzymes has been harnessed to activate tailored prodrugs within cancer cells that overexpress these enzymes, thus enhancing treatment specificity by enabling selective drug accumulation in a controlled manner tissue [26].

2 Materials and Methods

2.1 Software and tools

All analyses were performed on an Ubuntu 22.04 system. Sequence processing, alignment, tree generation, and database construction were performed using well-established bioinformatics tools with publicly available documentation; BLAST+ v2.16.0 [30], Bedtools v2.27.1 [2], Seqkit v2.3.0 [31], AliView v1.28 [14], Jalview v2.11.4.1 [12], MAFFT v7.526 [19], IQ-TREE v2.4.0 [13], and Figtree v1.4.4 [1]. All figures are placed at the end of the document in a separate Figures section.

Supplementary material is available at GitHub Repository.

2.2 Database and Query Sequence Preparation

To ensure comprehensive phylogenetic representation, thirteen primate genomes were selected from NCBI [21], spanning the major evolutionary clades: great apes, lesser apes, old world monkeys, new world monkeys, and prosimians (Table 2). *Homo sapiens* serves as the central reference point for comparison due to its well-characterized genome and key role in biomedical research. The great ape group includes *Pan troglodytes* (common chimpanzee), *Pan paniscus* (bonobo), *Gorilla gorilla* (western lowland gorilla), and two orangutan species, *Pongo pygmaeus* (Bornean orangutan) and *Pongo abelii* (Sumatran orangutan). Lesser apes are represented by *Hylobates moloch* (Javan gibbon), *Nomascus leucogenys* (northern white-cheeked gibbon), and *Symphalangus syndactylus* (siamang). Old world monkeys include *Macaca mulatta* (rhesus macaque), while New World monkeys are represented by *Callithrix jacchus* (common marmoset). The prosimian *Microcebus murinus* (gray mouse lemur) serves as an evolutionary outgroup, reflecting the most basal primate lineage. Each selected species has a publicly available, high-quality genome assembly, ensuring accurate homolog identification and supporting robust phylogenetic and comparative genomic analyses.

Three multi-gene families—GST, CYP, and CES were analysed across these primates. Known human members of each gene family were identified via the NCBI Gene database, and reference protein sequences were downloaded individually by searching gene names (e.g., *GSTA1 Homo sapiens*) in FASTA format. These sequences were concatenated into a single query dataset using the Unix *cat* command [35] to facilitate similarity searches.

Table 2. Genome assemblies used in the analysis

Species	Assembly
Homo sapiens (Human)	GRCh38.p14, T2T-CHM13v2.0
Pan troglodytes (Chimp-common)	NHGRI_mPanTro3-v2.1_pri
Pan paniscus (Chimp-Bonobo)	NHGRI_mPanPan1-v2.1_pri
Gorilla gorilla (Western lowland)	NHGRI_mGorGor1-v2.1_pri
Pongo pygmaeus (Bornean orangutan)	NHGRI_mPonPyg2-v2.1_pri
Pongo abelii (Sumatran orangutan)	NHGRI_mPonAbe1-v2.1_pri
Hylobates moloch (Javan gibbon)	Hmol_V3
Nomascus leucogenys (Northern white-cheeked gibbon)	Asia_NLE_v1
Symphalangus syndactylus (Siamang)	NHGRI_mSymSyn1-v2.1_pri
Callithrix jacchus (Marmoset)	mCalJa1.2.pat.X
Microcebus murinus (Gray mouse lemur)	Mmur_3.0
Macaca mulatta (Rhesus macaque)	Mmul_10

Protein databases for each species were constructed from their respective genome assemblies using the NCBI BLAST+ tool `makeblastdb`, with the `-parse_seqids` option to preserve original sequence identifiers for accurate tracking of BLAST hits. These steps were repeated for each primate genome so that databases for each primate and each gene family is constructed.

2.3 BLASTP and Postprocessing

Similarity searches were performed using `blastp` from the BLAST+ suite, querying the multi-gene family protein dataset against each species-specific protein database. Parameters were optimized to balance sensitivity and specificity: an E-value cutoff of 1e-5, word size of 2 to increase sensitivity for short alignments, and a maximum of 50 target sequences per query. The `-culling_limit 0` option ensured retention of all similar hits, including highly similar paralogs. Output was in tabular format (`-outfmt 6`) to facilitate downstream parsing.

BLASTP results were filtered using `awk` to remove low-confidence hits, retaining sequences with $\geq$20% identity and E-values $\leq 1e-5$. Unique sequence identifiers were extracted and used to retrieve corresponding sequences from the protein FASTA databases via Seqkit. Exact duplicate sequences were removed based on sequence content with `seqkit rmdup`. Subsequent manual curation was performed using Jalview to remove unrelated or spurious sequences.

2.4 Multiple sequence alignment

The curated multi-gene family protein sets were aligned separately using MAFFT v7.526. Initial alignments employed the E-INS-i algorithm with the `-maxiterate 1000` and `-genafpair` options, which are optimized for sequences containing multiple conserved domains and large insertions or deletions. Upon manual inspection and comparison of alignments generated using the `-localpair` and the `-globalpair` algorithms, it was observed that the local pair method produced fewer large gaps and better-conserved regions, thus the local pair algorithm was selected for the final alignments to maximize biological relevance.

2.5 Nomenclature

To maintain consistency and facilitate downstream analysis, a systematic gene nomenclature scheme was applied across all identified genes. Each gene was named using a species-specific prefix based on the Latin binomial abbreviation, followed by the gene symbol and isoform annotation. For example, genes from *Homo sapiens* were labeled with the prefix `Hsap`, followed by the standard gene symbol, such as `GSTA1`. Isoform variants were denoted by appending the isoform identifier as provided in the NCBI or UniProt database, e.g., `HsapGSTA1.X1` indicates isoform X1 of the *GSTA1* gene in human. Genes annotated as "like" or putative homologs received the suffix `.like`, for instance `HsapGSTA1.like`, to distinguish them from canonical gene models while reflecting their similarity. This nomenclature approach was consistently applied across all species by substituting the species prefix accordingly (e.g., `Ptro` for *Pan troglodytes*, `Mmul` for *Macaca mulatta*), allowing straightforward cross-species comparisons and clear identification of gene family members and isoforms.

The identification and classification of cytochrome P450 (CYP) genes presented several complications due to inconsistent annotation across genome assemblies. Although initial BLASTP searches retrieved numerous CYP homologs, many were labeled using functional enzyme names rather than standard CYP nomenclature. For example, CYP2R1 was annotated as "vitamin D 25-hydroxylase," CYP4F22 as "ultra-long-chain fatty acid omega-hydroxylase," and CYP5A1 as "thromboxane-A synthase." Other notable cases included steroidogenic enzymes such as CYP11A1 ("cholesterol side-chain cleavage enzyme"), CYP17A1 ("steroid 17-alpha-hydroxylase/17,20 lyase"), CYP19A1 ("aromatase"), and mitochondrial vitamin D–related enzymes such as CYP24A1 and CYP27B1. Additionally, some enzymes involved in sterol or cholesterol metabolism, such as CYP7A1, CYP7B1, CYP27A1, and CYP46A1, were annotated only by their substrate specificity (e.g., "cholesterol 7-alpha-monooxygenase" or "24-hydroxycholesterol 7-alpha-hydroxylase"). This necessitated manual curation to correctly map functionally annotated proteins to their corresponding CYP gene names. Such discrepancies underscore the need for consistent annotation standards across genome databases, especially when dealing with large multigene families like CYPs.

Table 3. Function-based annotation of CYP genes observed during BLASTP analysis. Several cytochrome P450 enzymes were named according to their biochemical activity rather than their canonical gene symbols, requiring manual curation.

CYP Gene Symbol	**Function-Based Name**
CYP2R1	Vitamin D 25-hydroxylase
CYP4F22	Ultra-long-chain fatty acid omega-hydroxylase
CYP5A1	Thromboxane-A synthase
CYP8A1	Prostacyclin synthase
CYP8B1	7-alpha-hydroxycholest-4-en-3-one 12-alpha-hydroxylase
CYP11A1	Cholesterol side-chain cleavage enzyme, mitochondrial
CYP17A1	Steroid 17-alpha-hydroxylase/17,20 lyase
CYP19A1	Aromatase
CYP21A2	Steroid 21-hydroxylase
CYP24A1	1,25-dihydroxyvitamin D(3) 24-hydroxylase, mitochondrial
CYP27A1	Sterol 26-hydroxylase, mitochondrial precursor
CYP27B1	25-hydroxyvitamin D-1 alpha hydroxylase, mitochondrial
CYP39A1	24-hydroxycholesterol 7-alpha-hydroxylase
CYP46A1	Cholesterol 24-hydroxylase
CYP51A1	Lanosterol 14-alpha demethylase
CYP7B1	25-hydroxycholesterol 7-alpha-hydroxylase
CYP7A1	Cholesterol 7-alpha-monooxygenase

2.6 Phylogenetic analysis

Phylogenetic trees for GST and CES were constructed using IQ-TREE v2.4.0 [13]; a fast and efficient software for maximum likelihood analysis with integrated model selection. The curated multiple sequence alignments served as input for IQ-TREE. The command `iqtree2 -s input.fasta -m MFP -B 1000 -T 8` was executed, where `-m MFP` directed IQ-TREE to perform ModelFinder to identify the best-fit substitution model where, `-B 1000` specified 1000 ultrafast bootstrap replicates to assess branch support, and `-T 8` enabled parallel processing using 8 CPU threads to accelerate computation. The resulting tree files were visualized and annotated using Figtree v1.4.4 [1], which allowed the application of customized color gradients to branches and clades to improve interpretability and aesthetic quality. Final trees were exported in high-resolution formats suitable for publication, ensuring clear visualization of phylogenetic relationships across the primate species analyzed.

Due to the large size of the CYP gene dataset (over 1200 sequences) and the computational intensity of performing maximum likelihood phylogenetic inference with IQ-TREE2, including 1000 ultrafast bootstrap replicates, the runtime exceeded the deadline on the available computational resources. To generate a phylogenetic tree within a reasonable timeframe, FastTree (version 2.1.11) was employed. FastTree uses an approximate maximum likelihood approach optimized for large alignments, balancing computational efficiency with reliable tree estimation [18]. The LG substitution model with gamma-distributed rate heterogeneity (-lg -gamma) was selected to model amino acid evolution. While FastTree provides an approximate solution and does not perform traditional bootstrapping, it is widely accepted in phylogenetics for large datasets where computational efficiency is essential [27].

3 Results

For the phylogenetic analysis of Glutathione S-Transferases (GSTs), a total of 13 primate genomes were included (Table 2). A total of 575 GST genes were identified across these genomes, including representatives from *Homo sapiens* (both GRCh38.p14 and T2T-CHM13v2.0), *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo pygmaeus*, *Pongo abelii*, *Symphalangus syndactylus*, *Macaca mulatta*, *Microcebus murinus*, *Callithrix jacchus*, *Nomascus leucogenys*, and *Hylobates moloch*.

Following the GST analysis, *H. sapiens* T2T-CHM13v2.0 and *H. moloch* (Hmol_V3) were excluded from subsequent carboxylesterase (CES) and cytochrome P450 (CYP) analyses. This decision was based on the extensive duplication of gene content between the GRCh38.p14 and T2T-CHM13v2.0 human assemblies, and the poor gene annotation quality observed in the Hmol_V3 assembly. Furthermore, the inclusion of two other gibbon genomes, *N. leucogenys* and *S. syndactylus*, is sufficient for the representation of the Hylobatidae family.

Subsequently, 272 CES genes and 1,285 CYP genes were identified across the remaining 11 primate genomes: *H. sapiens*, *P. troglodytes*, *P. paniscus*, *G. gorilla*, *P. pygmaeus*, *P. abelii*, *S. syndactylus*, *M. mulatta*, *M. murinus*, *C. jacchus*, and *N. leucogenys*.

3.1 Phylogenetic Analysis of CESs

Carboxylesterases (CES) comprise a diverse multigene enzyme family involved in the hydrolysis of ester-, amide-, and thioester-containing substrates, playing a key role in the metabolic processing of various xenobiotics and therapeutic agents. The carboxylesterase superfamily includes several characterized members: CES1, CES2, CES3, CES4, and CES5 Williams2010. Comparative analysis of *CES* gene family members across primates (Table 4) reveals variability in gene presence, copy number, and isoform diversity among species, reflecting complex lineage-specific patterns of gene expansion, loss, and diversification.

The observed differences in CES gene copy numbers across primate species suggest lineage-specific expansions that may indicate the evolutionary and functional relevance of particular CES gene families in each primate. For example, *Homo sapiens* shows a notable expansion in the *CES4* gene family, with 20 CES4 isoform entries, implying a possible divergence in function and an increased reliance on CES4-related activity in humans. Similarly, the chimpanzee (*Pan troglodytes*) exhibits a greater number of *CES1* gene variants (totaling 10) and *CES3* gene variants (totaling 8), suggesting these gene families may play a more prominent biological role in detoxification or lipid metabolism within this species. This pattern continues across other primates; for instance, the mouse lemur (Microcebus murinus) displays expansions in both CES1 (10 variants) and CES5 (7 variants), possibly indicating adaptations to distinct ecological niches or metabolic requirements. Such expansions may reflect selective pressures acting on CES gene subfamilies in different evolutionary lineages, where duplication and retention of certain CES genes likely confer an adaptive advantage. Therefore, the most expanded CES gene families in each primate can be interpreted as both more divergent and biologically significant within that species' genomic and physiological context. Gene duplicates may be retained through subfunctionalization or neofunctionalization, while others may degenerate into pseudogenes or evolve into lineage-specific isoforms [6]. The abundance of "-like" sequences, particularly in apes, suggests a complex interplay of these evolutionary processes, necessitating careful discrimination between functional isoforms and nonfunctional relics. For example, the canonical copies of CES4 and CES5 genes are often absent or present in low copy numbers across primates, yet numerous CES4-like and CES5-like sequences are observed. This pattern may reflect subfunctionalization or neofunctionalization of divergent paralogs following gene duplication events, consistent with models proposing that complementary loss of gene subfunctions can promote the preservation of duplicates [6]. Lemurs retain high counts of CES1-like, CES3 isoforms, and CES5 isoforms, suggesting that even when canonical CES genes are lost, isoform diversification can persist—potentially driven by lineage-specific selective pressures. This observation aligns with evidence that gene duplications may be adaptively maintained or lost in response to environmental conditions, highlighting the dynamic nature of gene family evolution [14].

The *CES1* canonical gene is rare among primates, present only in chimpanzees (*P. troglodytes*) and orangutans (*P. abelii*), each with one copy, suggesting either lineage-specific gene loss or annotation gaps. Humans (*H. sapiens*) lack a canonical *CES1* gene but have a relatively high number of *CES1* isoforms (5 copies) and one *CES1-like* gene identified. Every other primate in the analysis also has several isoforms apart from *M. murinus*, indicating possible paralog diversification across all primates. The higher numbers of isoforms across all primates are indicative of alternative isoform diversification, possibly driven by unique selective pressures.

For *CES2*, canonical genes are found in gorillas (*G. gorilla*), orangutans (*P. abelii*), gibbons (*N. leucogenys*), and macaques (*M. mulatta*), although only with a single copy, indicating these lineages may maintain conserved *CES2* functions. Interestingly, *S. syndactylus* has the highest number of *CES2* isoforms (5 copies), suggesting that isoform diversification may compensate for the absence of the canonical gene. The same pattern is observed in chimpanzees and humans, which lack the canonical *CES2* gene but have multiple isoforms, emphasizing isoform expansion as a compensatory mechanism.

Canonical *CES3* genes are notably absent in most species surveyed, with only gorilla (*G. gorilla*) and orangutan (*P. abelii*) retaining full-length *CES3* genes. This restricted presence may reflect either gene loss or incomplete genome annotations, particularly in poorly-annotated species such as Gibbons. Interestingly, the most *CES3 Isoforms* are found in *P. troglodytes* (10 genes), *P. pygmaeus* (13 genes), pointing to possible ape-specific paralog retention or subfunctionalization. Counts of 2-8 *CES3* genes are also observed across the other primates in the study, indicating ongoing diversification through alternative splicing or duplications of *CES3*-related sequences. The presence of isoform-like sequences, particularly the 7 copies in *P. pygmaeus*, may represent either partial duplicates, pseudogenes, or *CES3*-derived sequences that have diverged significantly in structure or function. This pattern implies that while full *CES3* gene structures may be rare or lost in many lineages, *CES3* enzymatic activity could be retained via non-canonical or truncated forms with potentially modified biochemical roles. Interestingly, no *CES3* were found in *M. mulatta* suggesting that *CES3* might have been retained in some apes but lost or altered in monkeys like M. mulatta.

CES1D, or carboxylesterase 1D, has been studied in mice, where its deficiency has been shown to protect against high-sucrose diet-induced hepatic triacylglycerol accumulation [17], modulate inflammatory responses in lung tissues [29], and potentially contribute to insulin resistance [4]. Despite these insights, the role of *CES1D* in primates remains unknown. Notably, the exclusive presence of *CES1D* in macaques suggests a recent duplication or retention event unique to Old World monkeys, which may indicate lineage-specific functional divergence within the CES gene family.

Overall, the variable distribution of CES genes across primates highlights the dynamic evolutionary pressures that shape their retention or loss in primates. These patterns likely correspond to species-specific metabolic adaptations, environmental challenges, or endogenous substrate specificities, reinforcing the notion that gene family evolution is tailored to the ecological and physiological context of each lineage.

It is important to acknowledge the limitations of using only BLASTP on gene models when interpreting these results. BLASTP searches against annotated proteomes; thus the analysis is restricted to previously characterized protein-coding sequences. The frequent presence of "-like" genes—homologous to known CES sequences—may reflect pseudogenes, partial gene fragments, or divergent paralogs without confirmed functionality. Consequently, some apparent gene absences could result from annotation gaps rather than gene loss. Zhang et al. estimated that nearly 50 percent of gene models in the rhesus macaque draft genome are missing, incomplete, or incorrect due to sequencing limitations and annotation errors [34]. Furthermore, even well-characterized genomes are not immune to such artifacts; Meyer et al. demonstrated that gene prediction errors persist in primate proteomes due to algorithmic constraints and inconsistencies across annotation pipelines [20].

The figure below (Figure 1) presents a rooted circular phylogenetic tree illustrating the evolutionary relationships among CES gene family members in primates. The tree is color-coded by gene type for clarity: CES4A (green), CES1 (red), CES2 (purple), CES5A (dark blue), and CES3 (light blue). The root is located at the center, with the CES1 cluster positioned closest to it. Branches radiate outward, forming distinct clusters that correspond clearly to each CES gene type, indicating strong sequence divergence among the gene families. Notably, the color groups largely remain separated, reflecting evolutionary divergence between CES gene families, with one exception where *PabeCES4A.like* and *PpygCES4A.like* group within the CES1 cluster, suggesting potential misannotation of these sequences. Branch lengths represent evolutionary distances; the shortest distances are observed between CES5A (dark blue) and CES2 (purple), implying a closer evolutionary relationship between these two gene types. Longer branch lengths separate the other clusters, indicating that CES5A and CES2 are more closely related to each other than to CES3 and CES4A.

Figure 2 presents a rectangular (linear) version of the same phylogenetic data depicted in Figure 1. As observed in the circular tree, CES1, CES2, CES3, CES4A, and CES5A genes each form well-defined clades, with CES1 genes located closest to the root. This reinforces the interpretation that CES1 represents the most basal group among the CES gene families and likely diverged first. Consistent with the circular tree, CES5A and CES2 genes remain more closely related to each other than to CES3 or CES4A. This implies that CES2 and CES3 share a more recent common ancestor than with CES4A or CES5A.

A noteworthy observation within the CES3 clade is where several CES4A genes from *Callithrix jacchus* are unexpectedly clustered. These include *CjacCES4A.X2.like*, *CjacCES4A.X3.like*, *CjacCES4A.X4*, *CjacCES4A.X10*, *CjacCES4A.X14*, *CjacCES4A.X17*, *CjacCES4A.X6*, *CjacCES4A.X9*, *CjacCES4A.X7*, *CjacCES4A.X8*, *CjacCES4A.X11*, *CjacCES4A.X1*, and *CjacCES4A.X5*. This unexpected grouping may indicate either misannotation of these sequences (i.e., they are CES3 paralogs) or lineage-specific gene duplication events in *C. jacchus*, where CES4A-like genes have diverged from CES3 ancestors but retain CES4A-like features.

It is also observed that, within each major CES gene clade, genes from *Callithrix jacchus* consistently cluster together into a small, distinct subclade positioned near the top of the clade, while *Microcebus murinus* genes form a separate subclade toward the bottom. This pattern is consistent across all CES gene families analyzed, where genes from each species cluster together in subclades. Such clustering is expected in species with longer independent evolutionary histories, such as *M. murinus* - the most basal primate in the dataset, as longer branch lengths and greater sequence divergence often reflect extended periods of lineage-specific evolution. The consistent grouping of genes from the same species or closely related species within individual CES gene clades strongly suggests lineage-specific divergence. Following the divergence of the major CES gene families, gene copies in each primate lineage appear to have evolved independently, gradually accumulating mutations over time. This pattern

Table 4: Distribution of CES gene family members and isoforms across primate species.

Gene Family	*H. sapiens*	*P. troglodytes*	*P. paniscus*	*G. gorilla*	*P. abelii*	*P. pygmaeus*	*N. leucogenys*	*S. syndactylus*	*C. jacchus*	*M. mulatta*	*M. murinus*
.	Hsap	Ptro	Ppan	Ggor	Pabe	Ppyg	Nleu	Ssyn	Cjac	Mmul	Mmur
CES1	-	1	-	-	1	-	-	-	-	-	-
CES1-like	1	1	1	1	2	1	-	-	-	1	4
CES1 Isoforms	5	3	4	4	-	2	4	4	3	6	-
CES1 Isoform-like	-	5	5	2	4	3	-	2	2	-	6
CES2	-	-	-	1	1	-	1	-	-	1	-
CES2-like	-	-	2	1	-	-	1	-	-	-	1
CES2 Isoforms	4	3	4	2	3	4	-	5	4	3	-
CES2 Isoform-like	-	1	-	2	-	-	-	1	-	-	-
CES3	-	-	-	1	1	-	-	-	-	-	-
CES3-like	-	3	-	1	1	-	-	-	-	-	-
CES3 Isoforms	3	5	8	3	1	6	5	5	2	-	6
CES3 Isoform-like	-	-	2	-	-	7	-	-	-	-	-
CES4	-	-	-	-	-	-	-	-	-	-	1
CES4-like	-	3	2	5	2	2	1	-	-	1	-
CES4 Isoforms	20	-	-	5	-	-	-	-	13	-	-
CES4 Isoform-like	-	-	-	-	-	-	-	-	2	-	-
CES5	-	-	1	-	1	1	-	-	-	1	-
CES5-like	-	-	-	1	-	-	-	1	1	1	-
CES5 Isoforms	3	4	-	-	-	-	2	-	-	3	7
CES5 Isoform-like	-	-	-	3	-	-	-	-	-	-	-
CES1D	-	-	-	-	-	-	-	-	-	1	-

of species-specific clustering highlights the evolutionary dynamics shaping the CES gene repertoire across primates.

3.2 Phylogenetic Analysis of GSTs

The glutathione S-transferase (GST) multigene family exhibits greater size and diversity compared to the carboxylesterase (CES) family, as demonstrated by the broader distribution of GST members across species (Table 5). The GST family consists of three superfamilies: the cytosolic, mitochondrial, and microsomal. Several enzymes, such as GST-kappa 1 (GSTK1), prostaglandin E synthase (PTGES), and the microsomal GSTs (MGST1, MGST2, MGST3), exhibit glutathione transferase-like activity but are not considered part of the evolutionary lineage of the cytosolic GST gene family [24]. These genes are classified under the membrane-associated proteins in eicosanoid and glutathione metabolism (MAPEG) family [11], therefore, they are excluded from this analysis. It is believed that the GST-like function of these membrane-bound enzymes likely arose through convergent evolution, rather than through divergence from a common ancestral GST gene. In contrast, true cytosolic GSTs are characterized by the presence of both GST-N and GST-C domains, which are essential for their enzymatic activity. While some unrelated proteins may contain one of these domains and display GST-like properties, they do not belong to the core GST family [24]. True cytosolic GSTs in humans comprise 16 genes classified into six major subclasses: alpha (*GSTA*), mu (*GSTM*), omega (*GSTO*), pi (*GSTP*), theta (*GSTT*), and zeta (*GSTZ*) [24]. The distribution and presence of GST genes exhibit substantial variability among primates, reflecting complex evolutionary dynamics.

Comparative genomic analysis across primates reveals substantial lineage-specific variation in both the retention of canonical GST genes and the expansion of isoforms, as observed in Table 5. Early-diverging primates such as *Microcebus murinus* and *Callithrix jacchus* retain canonical forms of several GST genes, including *GSTA1* and *GSTA5*, whereas more recently diverged species such as humans, chimpanzees, and orangutans predominantly exhibit numerous isoforms with few or no canonical genes. This pattern suggests that the ancestral GST genes were conserved in early-diverging primates, while in great apes, gene duplication events followed by divergence have led to a proliferation of isoforms. The apparent loss or divergence beyond recognition of canonical sequences in these species may reflect lineage-specific gene loss, functional specialization through subfunctionalization or neofunctionalization. Conversely, the pattern observed for *GSTA2* is opposite, with canonical genes present in humans and great apes but fewer isoforms compared to older primates. Canonical forms of *GSTA4* are conserved across most primate species, suggesting functional preservation throughout primate evolution.

The canonical *GSTT1* gene is notably absent in all examined primate species except *Macaca mulatta*, where it appears as a single copy. In contrast, numerous *GSTT1* isoforms are detected across most species, which may reflect either substantial isoform diversity or potential discrepancies in genome annotation. The *GSTT2* gene is largely absent, with humans exhibiting only two isoforms and a canonical *GSTT2* sequence identified solely in orangutans. This suggests a possible loss of *GSTT2* in several primates. Isoforms of *GSTT4* are abundant in many species except for earlier diverging primates such as *Microcebus murinus* and *Macaca mulatta*, which harbor one and two canonical *GSTT4* copies, respectively. The presence of *GSTT3* is minimal or absent across the species studied, indicating a possible reduced functional relevance or gene loss during primate evolution. Phylogenetic analysis (Figure 3) positions *GSTT4* as the closest to the ancestral *GSTT* gene cluster, followed by *GSTT2* and its paralog *GSTT2B*, then *GSTT3*, with *GSTT1* being the most distantly related. These distribution patterns highlight dynamic evolutionary processes, including gene duplication, loss, and diversification within the *GSTT* gene family among primates. Moreover, the varying abundance of isoforms suggests complex regulatory evolution or inconsistencies in gene model annotations across genome assemblies.

The figures below illustrate both a circular phylogenetic tree (Figure 3) and a rooted rectangular phylogenetic tree (Figure 4) constructed to visualize the evolutionary relationships among GST gene family members in primates. Gene types are color-coded for clarity: GSTT genes are represented by a gradient from light to dark blue, GSTA genes by a purple gradient, GSTM genes by an orange-to-red gradient, GSTP genes in pink, GSTO1 and GSTO2 in light and dark green, respectively, and GSTZ genes in gold.

The GSTA clade appears closest to the root, suggesting that it may represent the most ancestral lineage within the GST gene family. Notably, *CjacGSTA.X3.like*, *CjacGSTA4.X4.like*, *CjacGSTA4.X1.like*, *CjacGSTA4.X2.like*, and *MmurGSTA4.like* form a distinct clade located nearest the root, supporting this inference. If GSTA evolved first, it may have originally served foundational detoxification roles, such as

Table 5: Distribution of GST gene family members and isoforms across primate species.

Gene Family	*H. sapiens*	*P. troglodytes*	*P. paniscus*	*G. gorilla*	*P. abelii*	*P. pygmaeus*	*N. leucogenys*	*S. syndactylus*	*C. jacchus*	*M. mulatta*	*M. murinus*
	Hsap	Ptro	Ppan	Ggor	Pabe	Ppyg	Nleu	Ssyn	Cjac	Mmul	Mmur
GSTA1	-	-	-	-	2	1	-	1	1	1	-
GSTA1-like	-	-	-	-	2	2	-	1	-	1	-
GSTA1 Isoforms	3	2	2	2	-	-	2	-	-	-	-
GSTA1 Isoforms-like	-	-	1	-	-	-	-	-	-	-	2
GSTA2	1	1	1	1	1	-	-	-	-	1	-
GSTA2-like	-	-	-	1	-	-	-	-	-	-	1
GSTA2 isoform	-	-	-	-	-	2	-	-	-	1	-
GSTA3	-	-	1	-	1	-	-	1	-	1	-
GSTA3-like	-	1	1	2	-	-	-	-	-	-	2
GSTA3 isoform	3	3	-	2	2	2	2	-	2	2	-
GSTA3 Isoform-like	-	-	-	-	-	-	-	-	1	-	-
GSTA4	1	1	1	-	1	1	1	1	1	1	-
GSTA4-like	-	-	-	-	-	-	-	-	-	-	1
GSTA4 isoform	1	-	1	2	-	-	-	-	-	-	2
GSTA4 Isoform-like	-	-	-	-	-	-	-	-	3	-	-
GSTA5	1	1	1	1	-	-	1	1	-	1	-
GSTA5-like	-	-	-	-	1	1	-	-	-	-	-
GSTA5 isoform	-	-	-	-	-	3	-	-	-	1	-
GSTM1	-	1	1	1	1	-	-	-	-	1	-
GSTM1-like	-	-	2	2	4	2	-	-	-	-	-
GSTM1 Isoform	3	1	2	-	-	-	1	6	2	-	-
GSTM2	-	1	1	1	-	-	1	-	1	1	1
GSTM2-like	-	1	1	-	1	2	-	-	-	-	1
GSTM2 Isoform	2	1	-	1	-	-	-	2	-	-	-
GSTM3	1	1	1	1	1	1	1	-	1	1	1
GSTM3 Isoform	-	2	1	2	-	-	1	3	-	-	-
GSTM4	-	1	1	1	1	1	-	-	-	1	-
GSTM4-like	-	-	-	-	-	-	-	-	-	-	-
GSTM4 Isoform	5	3	1	-	1	-	3	6	-	4	2
GSTM5	1	1	1	1	1	1	1	-	1	1	-
GSTM5-like	-	-	-	-	-	-	-	1	3	-	-
GSTM5 Isoform	1	-	-	-	1	-	-	-	-	-	-
GSTP1	1	1	1	1	1	1	1	1	-	1	-
GSTP1 isoform	-	-	-	-	-	-	-	-	2	-	-
GSTZ1	-	-	-	-	-	-	1	-	-	-	-
GSTZ1 Isoform	5	6	4	4	7	5	-	4	5	3	3
GSTO1	-	1	-	-	-	-	1	-	-	1	1
GSTO1-like	-	1	1	-	-	-	-	-	-	-	-
GSTO1 Isoform	3	3	3	2	2	2	-	2	2	1	-
GST01 Isoform-like	-	-	-	2	2	2	2	2	-	-	-
GSTO2	-	-	-	1	1	1	1	-	-	1	1
GSTO2 Isoform	8	3	2	-	-	-	-	2	5	2	-
GSTT1	-	-	-	-	-	-	-	-	-	1	-
GSTT1-like	-	1	1	-	-	-	-	-	-	-	-
GSTT1 Isoform	6	3	4	5	4	3	5	6	5	6	2
GSTT1 Isoform-like	-	-	-	-	-	-	-	-	2	-	-
GSTT2-like	-	-	-	1	1	-	-	-	-	-	-
GSTT2 Isoform	2	-	-	-	1	-	-	-	-	-	-
GSTT2B	-	1	-	-	1	1	1	-	-	1	-
GSTT2B-like	-	-	-	-	1	1	-	-	-	-	1
GSTT2B Isoform	2	2	4	-	1	-	-	2	-	2	-
GSTT3	-	1	-	-	-	-	-	-	-	-	1
GSTT3-like	-	-	1	-	1	-	1	1	-	-	-
GSTT3 Isoforms	-	-	-	-	-	-	-	-	1	-	-
GSTT4	1	1	-	2	-	-	1	-	1	2	1
GSTT4-like	-	-	1	-	1	-	1	-	-	-	-
GSTT4 Isoform	6	9	3	-	5	3	-	2	3	-	-
GSTT4 Isoform-like	-	-	-	2	5	4	-	2	5	-	-

conjugation of glutathione to xenobiotics or endogenous metabolites, then later-evolving subfamilies (e.g., GSTM, GSTP, GSTO, GSTZ) may have diverged to take on more specialized roles or adapted to different substrate specificities.

Within the phylogenetic tree, the GSTM and GSTP genes are observed to cluster more closely together, forming adjacent clades. Similarly, GSTO and GSTZ genes are located in close proximity to one another. This suggests that each pair—GSTM-GSTP and GSTO-GSTZ—may share a more recent common ancestor compared to other GST subfamilies. Similar to patterns observed in the CES gene family, the GST gene tree shows clear color-coded clustering corresponding to distinct gene types, reinforcing the evolutionary distinctness of each GST subfamily. Also consistent with CES observations, genes from *Callithrix jacchus* and *Microcebus murinus* often form more distantly placed subclades compared to those from other primates. For example, at the top of the rooted tree (Figure 4), *CjacGSTA.X3.like*, *CjacGSTA4.X4.like*, *CjacGSTA4.X1.like*, and *CjacGSTA4.X2.like* form a species-specific subclade, with *MmurGSTA4.like* located nearby in a neighboring branch. Toward the bottom of the tree, a distinct clade includes *MmurGSTA3.like_1*, *MmurGST.X1.like*, *MmurGST.X2.like*, *PabeGST.like*, and *PpygGST.like*, while another nearby clade contains *CjacGSTA4*, *MmulGSTA4*, *NleuGSTA4*, *PtroGSTA4*, *GgorGSTA4.X2*, *Hsap2GSTA4.X2*, *MmurGSTA4.X1*, and *MmurGSTA4.X2*.

3.3 Phylogenetic Analysis of CYPs

The cytochrome P450 (CYP) superfamily comprises 18 distinct mammalian families, collectively encoding 57 genes in the human genome [23, 5]. These CYP enzymes play a central role in the metabolism of a wide range of drugs, with families CYP2, CYP3, and CYP4 containing the largest number of genes [23]. Among these, CYP3A4, CYP2C9, CYP2C8, CYP2E1, and CYP1A2 are predominantly expressed in the liver, making them key contributors to hepatic drug metabolism. Other isoforms such as CYP2A6, CYP2D6, CYP2B6, CYP2C19, and CYP3A5 are present at lower levels in the liver, while certain CYP enzymes, including CYP2J2, CYP1A1, and CYP1B1, are mainly found in extrahepatic tissues [33].

Table 6. Classification of CYP genes by functional groups

Functional Group	Genes
Xenobiotics	2C8, 1A1, 3A7, 2E1, 2A2, 2D6, 3A4
	2B6, 1A2, 2F1, 3A5, 2C9, 2A13, 2C19, 2C18
Sterols	17A1, 11A1, 11B2, 21A2, 51A1, 27A1, 7A1
	19A1, 8B1, 39A1, 46A1, 11B1, 1B1, 7B1
Unknown	4A22, 2S1, 20A1, 2A7, 2W1
Fatty acids	4Z1, 2U1, 4B1, 4F12, 4F22, 4X1, 2J2
	4F11, 4A11, 4V2
Eicosanoids	4F2, 5A1, 8A1, 4F8, 4F3
Vitamins	26C1, 24A1, 26B1, 27B1, 26A1, 2R1

Over the past decade, there has been a substantial expansion in the number of identified cytochrome P450 (CYP) isozymes, with more than 300,000 distinct CYP proteins identified by 2018 [10], and this trend is expected to continue as more genomes are sequenced. Despite their diversity, all known CYP enzymes share a conserved region including: a P450 signature motif as well as other conserved regions such as a tetrapeptide in the K helix, an aromatic region between helix K and the P450 signature sequence, and a pentapeptide sequence in the C helix of eukaryotic P450s [16].

Figure 5 displays a circular phylogenetic tree illustrating the evolutionary relationships among the cytochrome P450 (CYP) gene family. This analysis encompasses a comprehensive dataset comprising over 1200 CYP gene sequences identified across 11 primate species, highlighting the extensive diversification within this gene family. Due to the large size and complexity of the dataset, the detailed rooted phylogenetic tree is partitioned into four pages. These sections are presented in Figures 6, 7, 8, and 9. Due to Fasttrees' inaccuracy with branch lengths, it is difficult to infer exact evolutionary trajectories. Despite this clear clustering of major CYP subfamilies is observed. Notably, the CYP2 family appears closest to the inferred root with the most genes found. This expansion of CYP2 indicates its biological relevance in primates. However, evidence has put CYP51 as one of the most ancient and universally conserved CYP families across eukaryotes, with it being present even in the earliest Eukaryote, followed by CYP61 and CYP710 [25].

Research on the evolution of CYP genes in primates is limited, but some studies have offered valuable insights. For example, investigations into the CYP2D subfamily revealed that the human CYP2D subfamily includes one functional gene, CYP2D6, and two nonfunctional paralogs, CYP2D7 and CYP2D8. CYP2D6 plays a key role in metabolizing alkaloids and approximately 25 percent of commonly used drugs. Comparative studies in primates suggest that CYP2D7 arose from a duplication event in a shared ancestor of humans and great apes, while the origins of CYP2D6 and CYP2D8 trace back to a common ancestor of New World monkeys and Catarrhini [32].

Despite the insights gained from this analysis, several limitations hinder a more comprehensive understanding of CYP gene evolution in primates. First, the use of BLASTP alone may have overlooked fragmented or pseudogenic sequences. Second, due to the large number of CYP genes and the computational demands associated with high-quality alignments and tree inference, the analysis of CYPs was constrained in both depth and scope. To improve upon this, future analyses could incorporate nucleotide data. Running a more robust maximum likelihood method, such as IQ-TREE with appropriate models and bootstrapping (like what was done with GSTs and CES), would also enhance the accuracy of evolutionary inference and branch length estimation. Collectively, these refinements would offer a more complete picture of CYP gene family diversification across primates.

4 Conclusion

This study demonstrates that the CES, GST, and CYP gene families have undergone extensive lineage-specific diversification across primates, driven by gene duplication, functional divergence, and differential retention of canonical gene forms. The observed interspecies variation in gene copy number and isoform composition underscores the evolutionary flexibility of these detoxification-related families. However, persistent annotation gaps and the presence of "-like" variants highlight important limitations of relying solely on annotated proteomes. Using BLASTP alone likely underestimates gene diversity due to incomplete gene models and inconsistent or incorrect annotations across genome assemblies. Additionally, pseudogenes were excluded from this analysis due to time constraints, which further limits the scope. These challenges reflect broader issues faced in comparative genomics when working with publicly available datasets. To overcome these limitations, future studies should incorporate nucleotide-level analyses, such as BLASTN-based searches, to uncover unannotated or misannotated genes and pseudogenes, thereby refining gene family characterizations and evolutionary insights.

Acknowledgements

I would like to sincerely thank Dr. Lars Jermiin for his guidance and encouragement throughout this research. I am also grateful to the National University of Ireland, Galway, for providing the excellent academic environment and support during my Master's program in Biomedical Genomics.

Figures

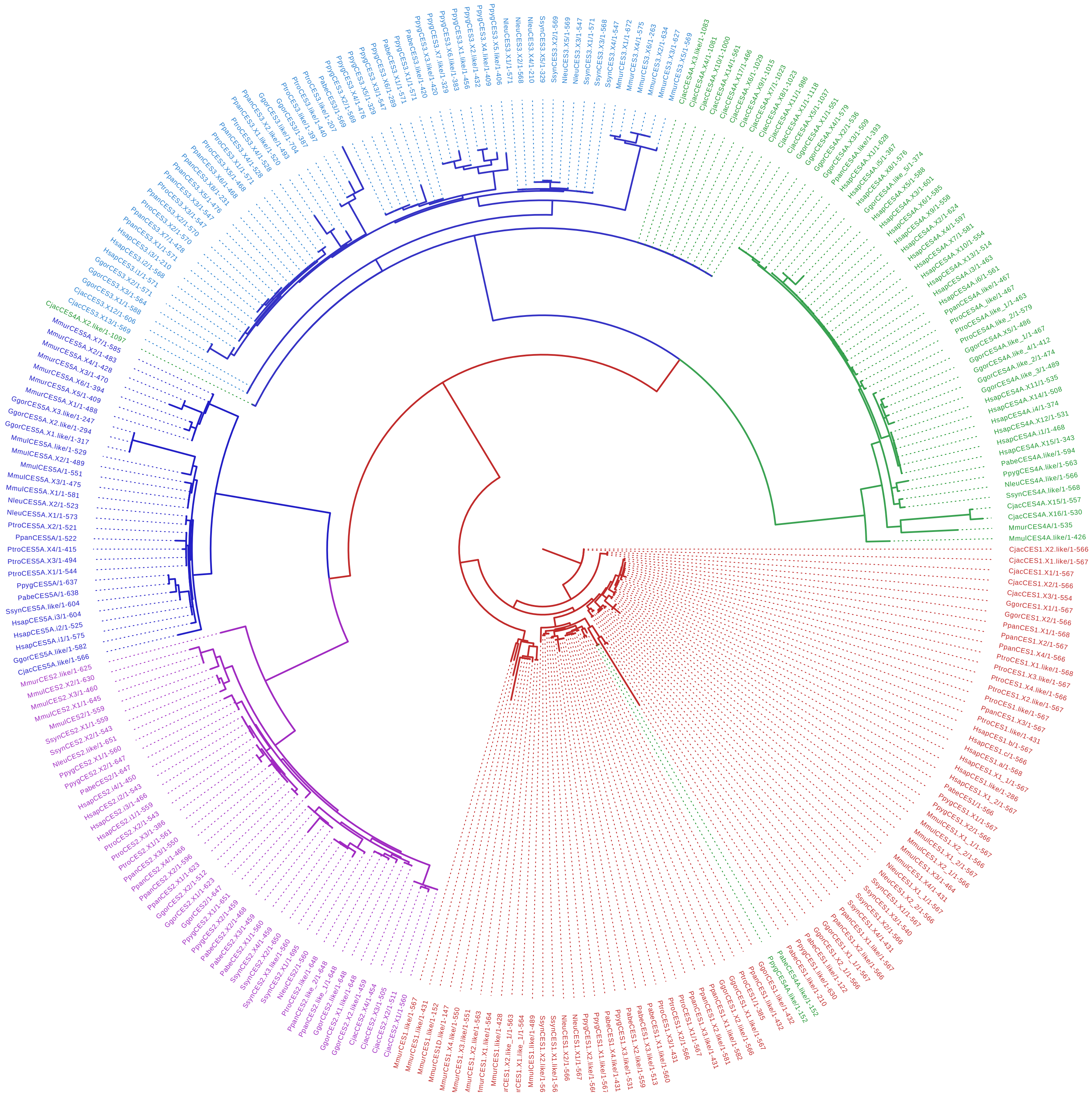

Figure 1: Circular illustration of the phylogenetic tree of CES genes in primates. The tree was inferred using IQ-TREE version 2.4.0 under the best-fit substitution model selected by ModelFinder. The circular layout facilitates visualization of evolutionary relationships among CES gene family members across multiple primate species.

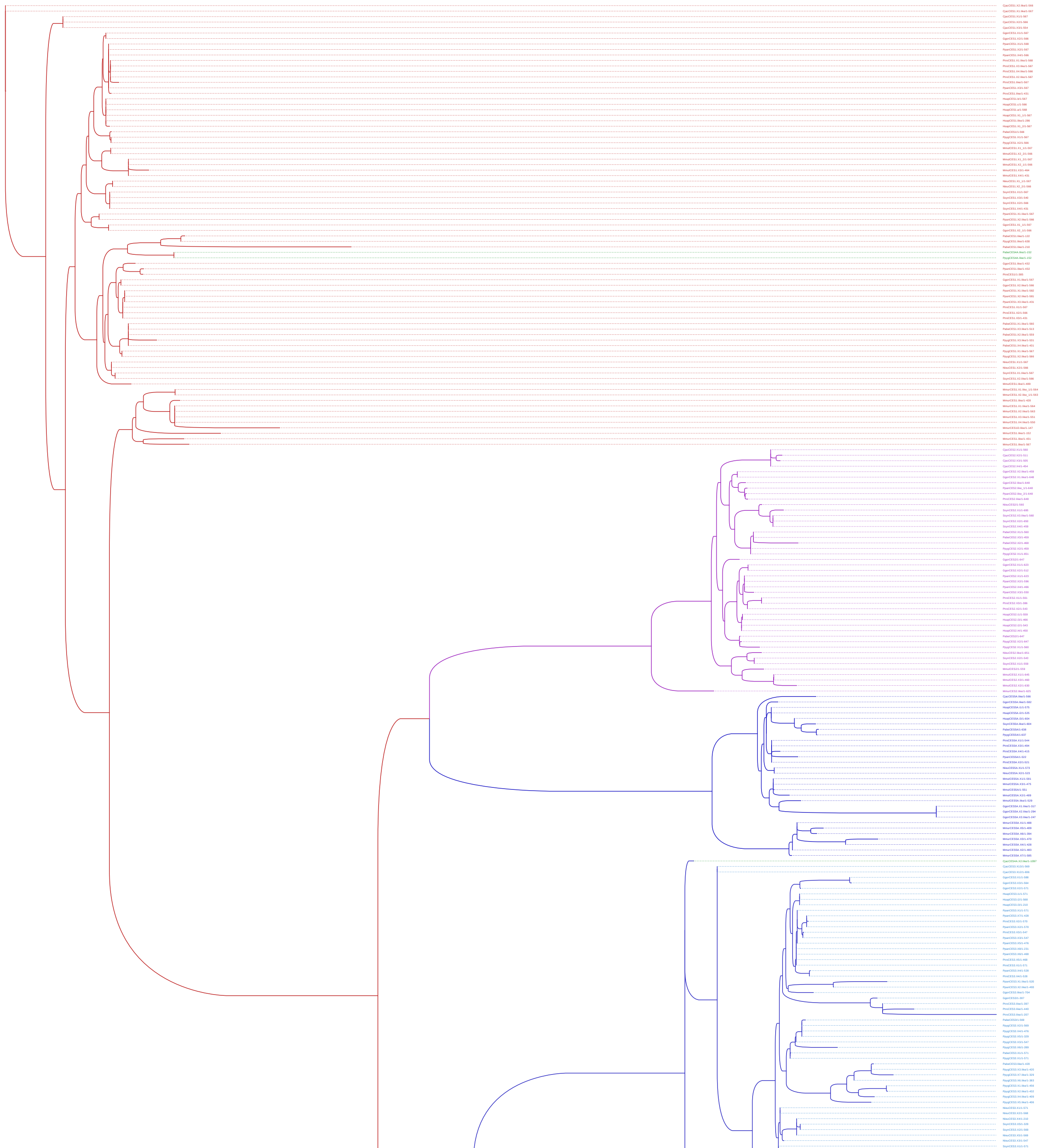

Figure 2: Rooted Illustration of phylogenetic tree of CES genes in primates, using IQ-TREE version 2.4.0.

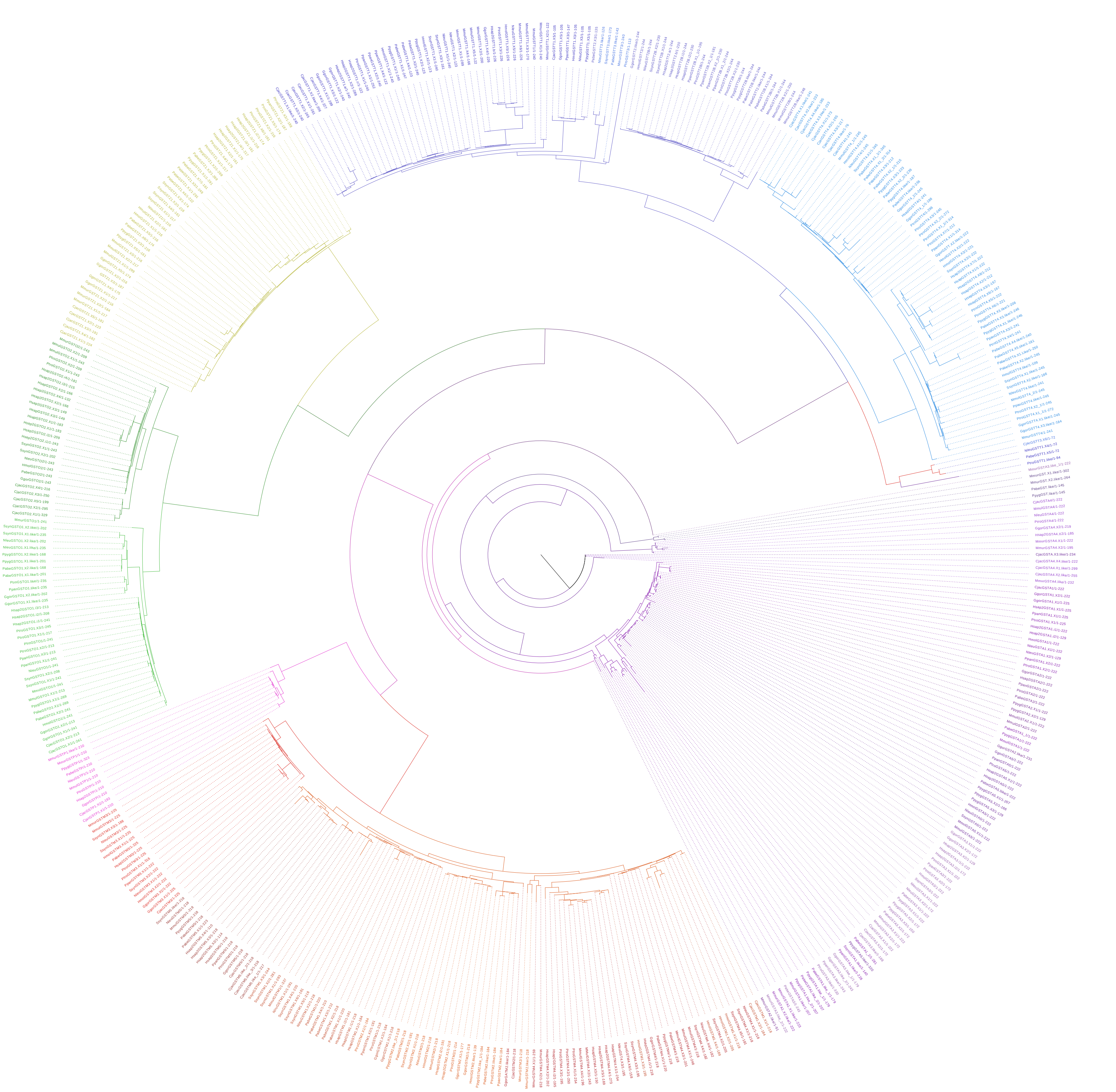

Figure 3: Circular illustration of the phylogenetic tree of GST genes in primates generated using IQ-TREE version 2.4.0

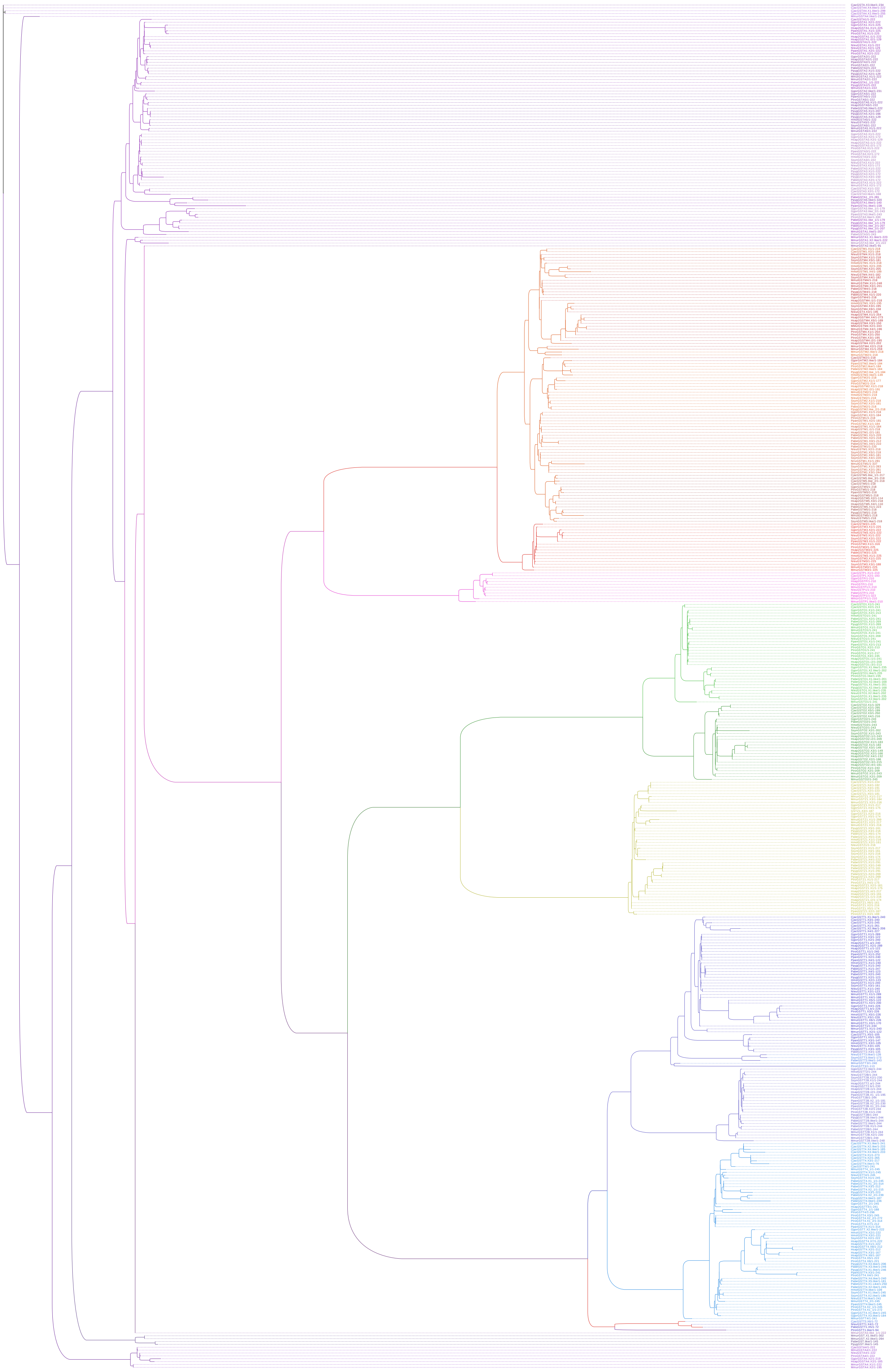

Figure 4: Rooted Illustration of phylogenetic tree of CES genes in primates, using IQ-TREE version 2.4.0.

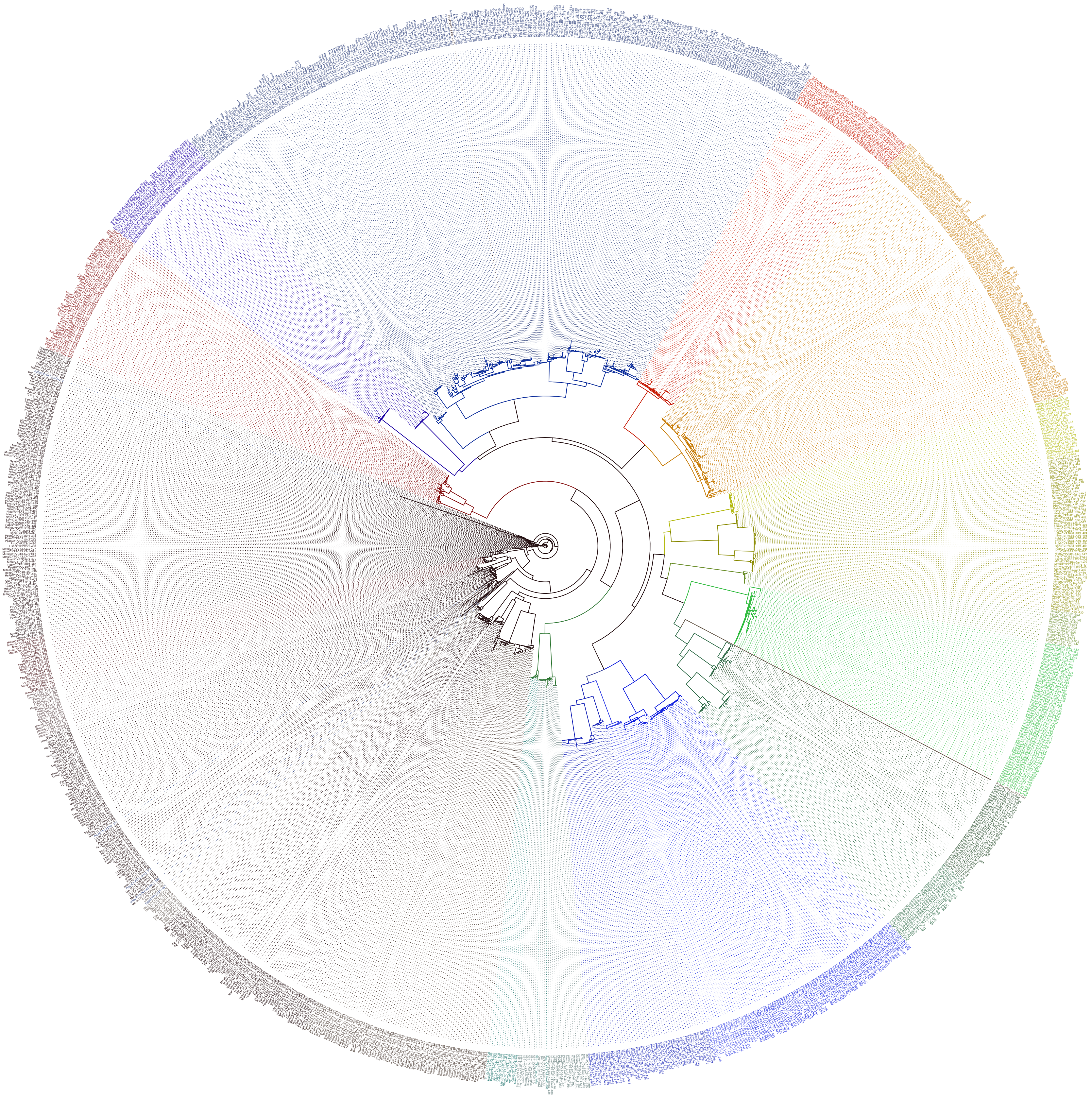

Figure 5: Circular Illustration of phylogenetic tree of CYP genes in primates, generated using FastTree v2.1.11.

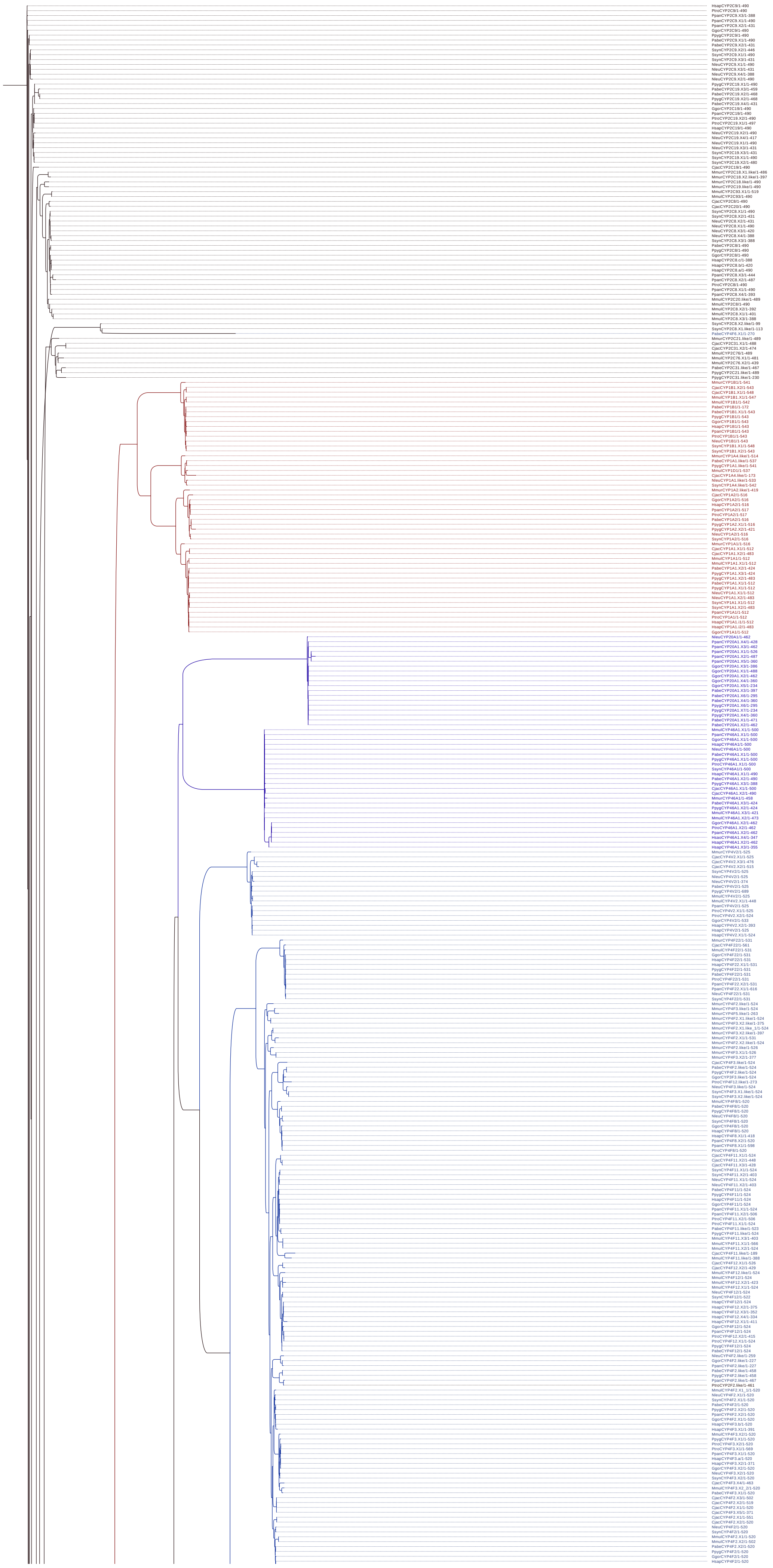

Figure 6: Rooted illustration of phylogenetic tree of CYP genes in primates (Part 1 of 4).

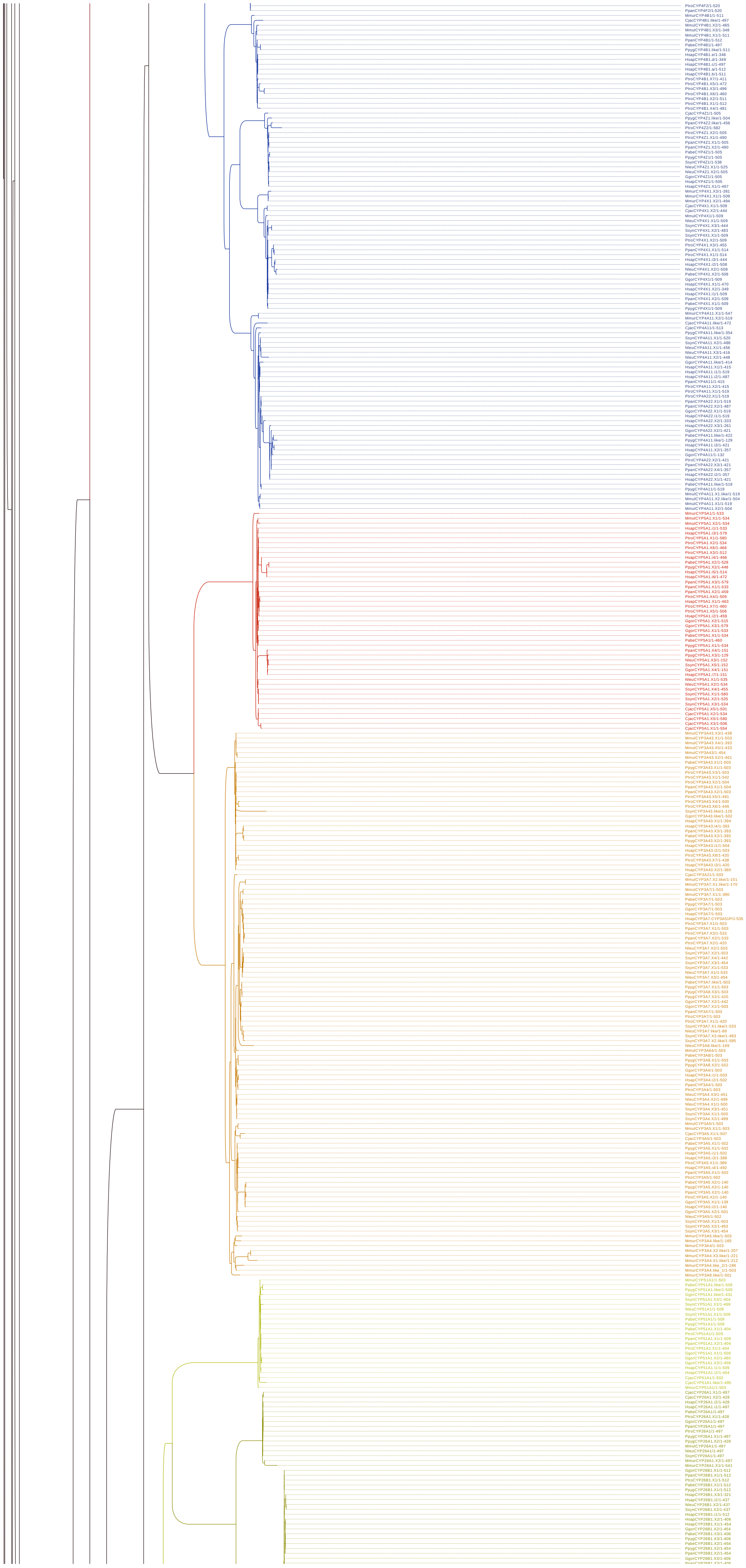

Figure 7: Rooted illustration of phylogenetic tree of CYP genes in primates (Part 2 of 4).

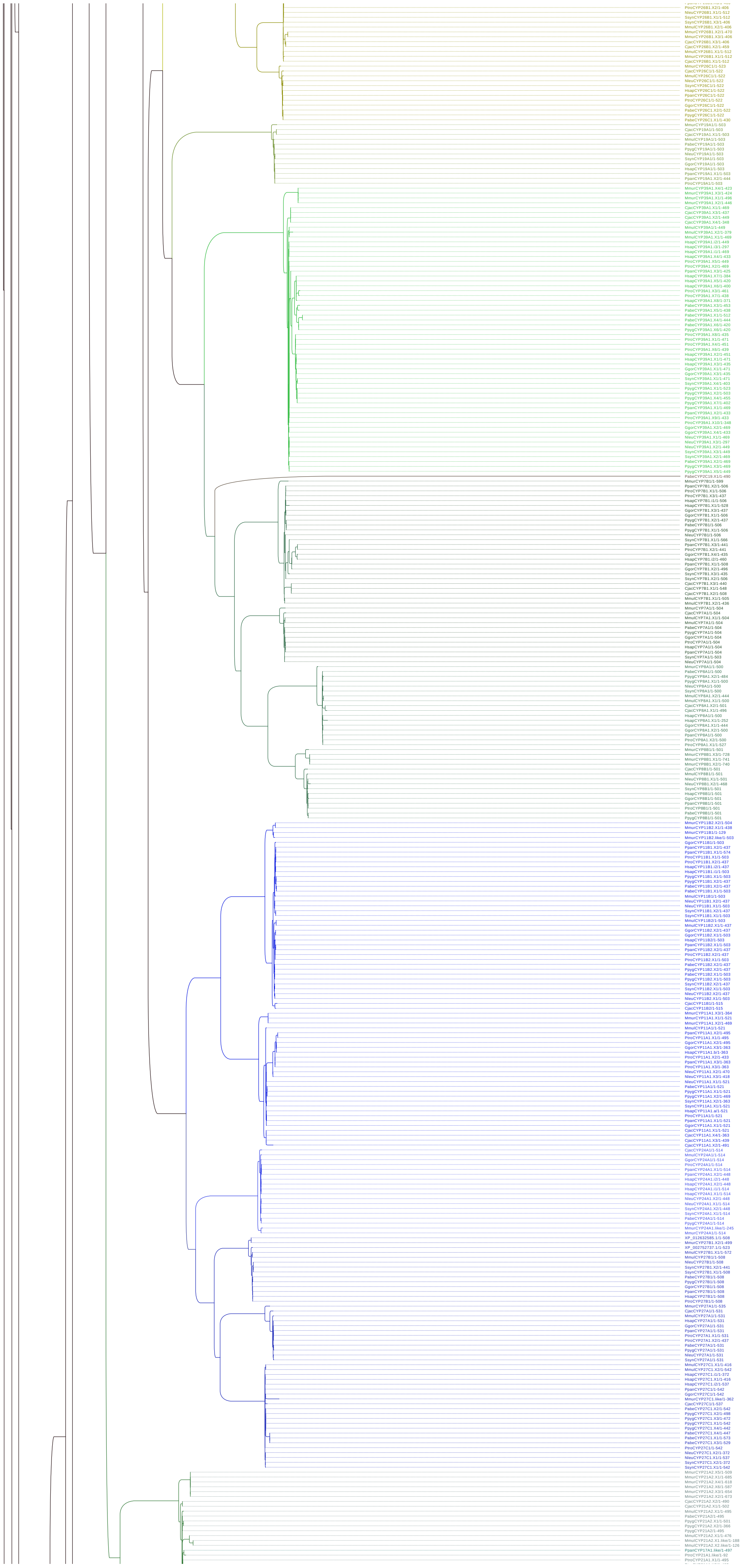

Figure 8: Rooted illustration of phylogenetic tree of CYP genes in primates (Part 3 of 4).

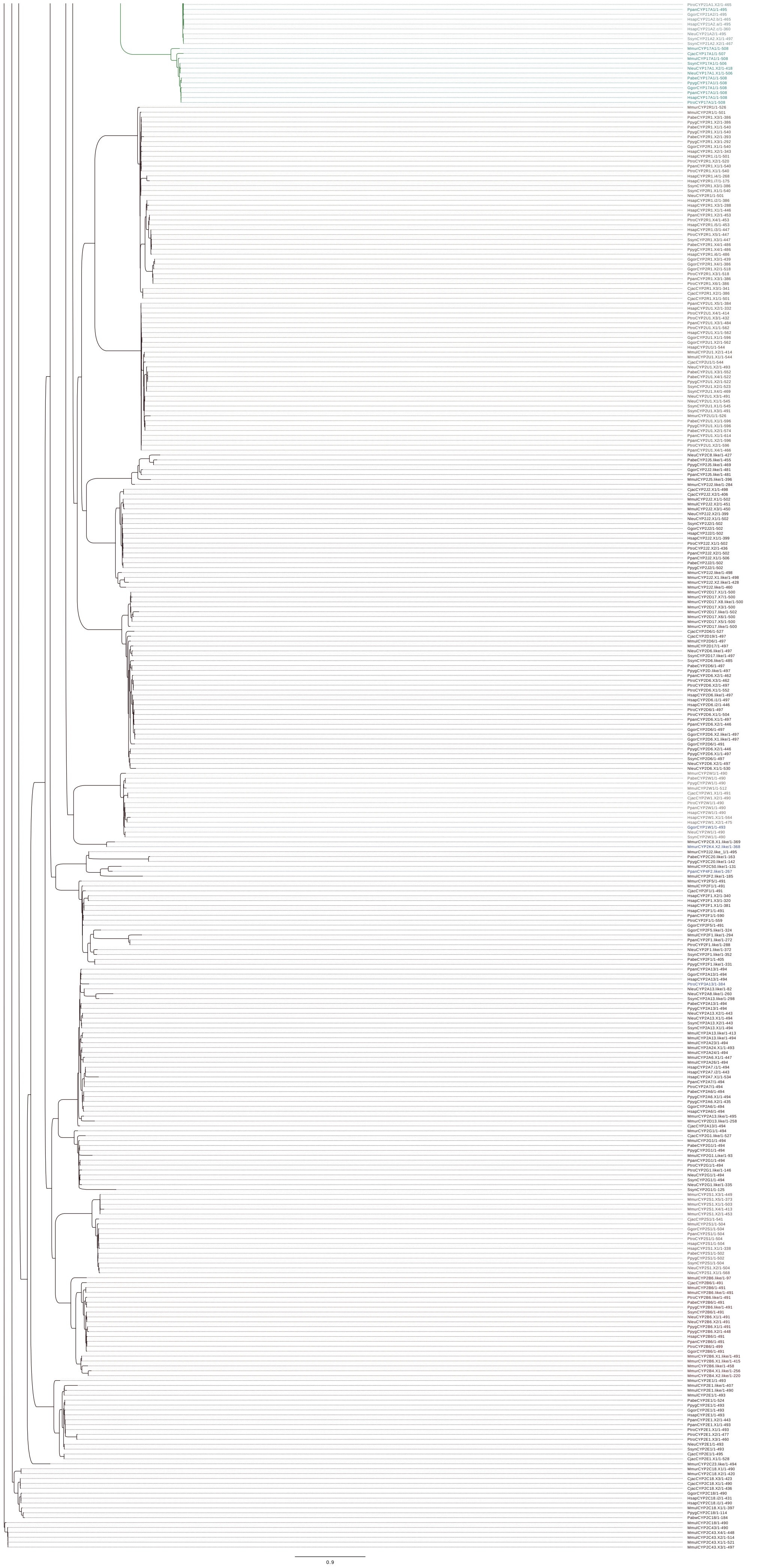

Figure 9: Rooted illustration of phylogenetic tree of CYP genes in primates (Part 4 of 4).

References

[1]Anaconda.org. Figtree | Anaconda.org. https://anaconda.org/bioconda/figtree, 2025. [Accessed 14 Jul. 2025].

[2]BEDTools Documentation. *getfasta — bedtools 2.31.0 documentation*, n.d. Accessed: 2025-07-14.

[3]S. Casey Laizure, V. Herring, Z. Hu, K. Witbrodt, and R.B. Parker. The role of human carboxylesterases in drug metabolism: Have we overlooked their importance? *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 33(2):210–222, 2013.

[4]R. Chen, Y. Wang, R. Ning, J. Hu, W. Liu, J. Xiong, L. Wu, J. Liu, G. Hu, and J. Yang. Decreased carboxylesterases expression and hydrolytic activity in type 2 diabetic mice through Akt/mTOR/HIF-1/Stra13 pathway. *Xenobiotica*, 45(9):782–793, 2015.

[5]F. Esteves, J. Rueff, and M. Kranendonk. The central role of cytochrome p450 in xenobiotic metabolism—a brief review on a fascinating enzyme family. *Journal of Xenobiotics*, 11(3):94–114, 2021.

[6]Bilal Gilani and Michele Cassagnol. Biochemistry, cytochrome p450, 2023. Accessed: 2025-07-23.

[7]F. P. Guengerich. Cytochrome p450 enzymes as drug targets in human disease. *Drug Metabolism and Disposition*, 52(6):493–497, 2023.

[8]Lu Her and Hao-Jie Zhu. Carboxylesterase 1 and precision pharmacotherapy: Pharmacogenetics and nongenetic regulators. *Drug Metabolism and Disposition*, 48(3):230–244, 2019.

[9]Roger S. Holmes, Jennifer Chan, Lisa A. Cox, William J. Murphy, and John L. VandeBerg. Opossum carboxylesterases: sequences, phylogeny and evidence for ces gene duplication events predating the marsupial-eutherian common ancestor. *BMC Evolutionary Biology*, 8(1):54–54, 2008.

[10]Magnus Ingelman-Sundberg. Cytochrome p450 polymorphism: From evolution to clinical use. *Advances in Pharmacology*, pp. 393–416, 2022.

[11]P.-J. Jakobsson, R. Morgenstern, J. Mancini, A. Ford-Hutchinson, and B. Persson. Membrane-associated proteins in eicosanoid and glutathione metabolism (mapeg). *American Journal of Respiratory and Critical Care Medicine*, 161(supplement_1):S20–S24, 2000.

[12]Jalview Development Team. Download - jalview. https://www.jalview.org/download/, n.d. Accessed: 2025-07-14.

[13]Subha Kalyaanamoorthy, Bui Quang Minh, Tung Kai Fung Wong, Arndt von Haeseler, and Lars S. Jermiin. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589, 2017.

[14]Anders Larsson. Aliview: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, 2014.

[15]D.F.V. Lewis and Y. Ito. Cytochrome p450 structure and function: An evolutionary perspective. In *Issues in Toxicology*, pp. 3–45. Royal Society of Chemistry, 2008.

[16]D.F.V. Lewis, E. Watson, and B.G. Lake. Evolution of the cytochrome p450 superfamily: sequence alignments and pharmacogenetics. *Mutation Research/Reviews in Mutation Research*, 410(3):245–270, 1998.

[17]J. Lian, R. Watts, A. D. Quiroga, M. R. Beggs, R. T. Alexander, and R. Lehner. Ces1d deficiency protects against high-sucrose diet-induced hepatic triacylglycerol accumulation. *Journal of Lipid Research*, 60(4):880–891, 2019.

[18]Kang Liu, Craig R. Linder, and Tandy Warnow. Raxml and fasttree: Comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE*, 6(11):e27731, 2011.

[19]MAFFT Development Team. Mafft - a multiple sequence alignment program. https://mafft.cbrc.jp/alignment/software/, n.d. Accessed: 2025-07-14.

[20]C. Meyer, N. Scalzitti, A. Jeannin-Girardon, P. Collet, O. Poch, and J.D. Thompson. Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. *BMC Bioinformatics*, 21(1):43, 2020.

[21]NCBI. Genome. https://www.ncbi.nlm.nih.gov/datasets/genome/, n.d. Accessed: 2025-07-14.

[22]D. W. Nebert, D. R. Nelson, and R. Feyereisen. Evolution of the cytochrome p450 genes. *Xenobiotica*, 19(10):1149–1160, 1989.

[23]D. W. Nebert, K. Wikvall, and W. L. Miller. Human cytochromes p450 in health and disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1612):20120431, 2013.

[24]D.W. Nebert and V. Vasiliou. Analysis of the glutathione s-transferase (gst) gene family. *Human Genomics*, 1(6):460, 2004.

[25]David R. Nelson. Cytochrome p450 diversity in the tree of life. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1866(1):141–154, 2018.

[26]Milica Pljesa-Ercegovac, Aleksandra Savic-Radojevic, Milica Matic, Vladana Coric, Tijana Djukic, Tatjana Radic, and Tatjana Simic. Glutathione transferases: Potential targets to overcome chemoresistance in solid tumors. *International Journal of Molecular Sciences*, 19(12):3785, 2018.

[27]Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, 2010.

[28]D. Sheehan, G. Meade, V. M. Foley, and C. A. Dowd. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochemical Journal*, 360(1):1, 2001.

[29]B. N. Szafran, Abdolsamad Borazjani, H. L. Scheaffer, J. A. Crow, A. M. McBride, O. Adekanye, C. B. Wonnacott, R. Lehner, and M. K. Ross. Carboxylesterase 1d Inactivation Augments Lung Inflammation in Mice. *ACS Pharmacology & Translational Science*, 5(10):919–931, 2022.

[30]Tao Tao. Standalone blast setup for unix. https://www.ncbi.nlm.nih.gov/books/NBK52640/, 2020. Accessed: 2025-07-14.

[31]Shenwei Wang. Seqkit - ultrafast fasta/q kit. https://bioinf.shenwei.me/seqkit/, n.d. Accessed: 2025-07-14.

[32]Yoshiki Yasukochi and Yoko Satta. Molecular evolution of the CYP2D subfamily in primates: Purifying selection on substrate recognition sites without the frequent or long-tract gene conversion. *Genome Biology and Evolution*, 7(4):1053–1067, 2015.

[33]Ute M. Zanger and Markus Schwab. Cytochrome p450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, 2013.

[34]X. Zhang, J. Goodsell, and R.B. Norgren. Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics*, 13(1):206, 2012.

[35]S. Zivanov. Cat command in linux {15 Commands with Examples} | phoenixnap kb. https://phoenixnap.com/kb/linux-cat-command, 2020. Accessed: 2025-07-14.