Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cell-Level Auditing of an Orthology-Database-Derived Gene Presence/Absence Matrix Reveals Lineage-Specific Gene Gain, Duplication, and Pseudogenization

Yu Ling

lingyu@imau.edu.cn

Inner Mongolia Agricultural University

Heping Fu

Inner Mongolia Agricultural University

Research Article

Keywords: OrthoDB, presence/absence matrix, cell-level auditing, competitive paralog discrimination, GLUD2 retrocopy, candidate GLUD1 duplication, UOX pseudogenization

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11314390/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cell-Level Auditing of an Orthology-Database-Derived Gene Presence/Absence Matrix Reveals Lineage-Specific Gene Gain, Duplication, and Pseudogenization

Yu Ling[1,2,3*] and Heping Fu[1,2,3,#]

[1] College of Grassland Science, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, China

[2] Inner Mongolia Autonomous Region Field Scientific Observation and Research Station for Desert Ecosystem and Rodent Pest Monitoring and Control, Alxa Left Banner, Inner Mongolia, China

[3] Key Laboratory of Grassland Resources, Ministry of Education, Hohhot, Inner Mongolia, China

* Corresponding author: lingyu@imau.edu.cn

Lead corresponding author: fuheping@126.com

Abstract

Cross-species gene presence/absence (PA) matrices are widely used to infer metabolic-pathway evolution, but orthology database clusters may contain deeply diverged homologs, lineage-specific duplicates, or pseudogenes, leading to erroneous assignments. Using OrthoDB v12 as the initial data source, we constructed a PA matrix comprising 34 human genes involved in nitrogen waste excretion and ammonia metabolism across 60 species, totaling 2,040 gene-by-species cells. Cells with uncertain assignments or distributions inconsistent with biological expectations were iteratively audited using NCBI/RefSeq searches, whole-genome tblastn, competitive paralog discrimination, locus-level clustering, phylogenetic priors, and published evidence.

The final matrix contained 412 cells without support for the focal ortholog, 1,578 OrthoDB-only presence calls, 38 presence calls additionally supported by RefSeq or tblastn, 6 ancestral or unresolved homolog calls, 5 known pseudogenes, and 1 cluster bleed-through call. Sequence-level

assessment of 20 randomly sampled OrthoDB-only presence cells yielded confirmation rates of 75% under a strict single-HSP criterion and 85% under a locus-aggregate criterion. Direct sequence evidence supported all four highest-risk absence assignments.

All six candidate GLUD loci examined in non-ape vertebrates favored GLUD1 over GLUD2, with five exceeding the 2σ discrimination threshold, supporting the ape-specific retrocopy origin of GLUD2. Two loci on different chromosomes in Anolis carolinensis showed GLUD1-favored signals, suggesting a candidate GLUD1 duplication in this lineage. UOX sequences from four great apes shared a codon 33 stop mutation, supporting UOX pseudogenization in the Hominidae ancestor. These findings demonstrate that database cluster membership cannot be equated directly with focal-ortholog or functional-gene presence. Cell-level auditing provides a transparent and traceable framework for interpreting orthology-database-derived PA matrices and lineage-specific gene gain, duplication, and pseudogenization.

Keywords: OrthoDB; presence/absence matrix; cell-level auditing; competitive paralog discrimination; GLUD2 retrocopy; candidate GLUD1 duplication; UOX pseudogenization

1. Introduction

Gene retention, duplication, and loss provide an important genetic basis for the evolution of metabolic capacities. Cross-species gene presence/absence matrices are widely used to infer the origins, assembly, and lineage-specific changes of metabolic pathways (Dessimoz, et al. 2012; Sonnhammer, et al. 2014). These matrices are commonly constructed from orthologous groups in databases such as OrthoDB, OMA, and eggNOG (Altenhoff, et al. 2024; Hernandez-Plaza, et al. 2023; Kuznetsov, et al. 2023). In downstream analyses, database cluster membership is often interpreted as gene presence, whereas the absence of a corresponding cluster member is interpreted as gene absence. Genome assembly quality, gene annotation, sequence divergence, duplication, and

database cluster boundaries may all affect these assignments. Database output therefore cannot be equated directly with focal-ortholog presence or absence.

Database-derived PA matrices may contain both false-positive and false-negative calls. A false positive may arise when another paralog in the target species produces a strong sequence-similarity signal and is misidentified as the focal ortholog. Alternatively, a deeply diverged homolog or ancient paralog may be grouped in the same database cluster as a recently derived lineage-specific gene. We refer to the latter case as cluster bleed-through. False-negative calls may result from assembly gaps, missing annotations, fragmented gene models, extensive sequence divergence, or inappropriate paralog grouping. Because sequence similarity alone generally cannot distinguish a focal ortholog from a paralog or nonfunctional remnant, PA calls require cell-level auditing using genomic searches, competitive paralog discrimination, phylogenetic information, and published evidence.

Genes involved in nitrogen assimilation, transport, detoxification, and excretion provide a suitable system for evaluating PA-matrix reliability because they include deeply conserved enzymes, lineage-specific paralogs, and pseudogenized loci. GLUD1 and GLUD2 offer a particularly informative test. GLUD2 originated through retrotransposition of a GLUD1 transcript in the ape ancestor approximately 18 to 25 Ma (Burki and Kaessmann 2004; Dimovasili, et al. 2021; Litso, et al. 2023; Plaitakis, et al. 2017; Rosso, et al. 2008). GLUD2-associated cluster members in non-ape species must therefore be distinguished from GLUD1, independently derived GLUD1-like duplicates, and other GDH-family homologs.

UOX provides a complementary test involving pseudogenization. A shared nonsense mutation at codon 33 supports UOX pseudogenization in the common ancestor of Hominidae, whereas Old World monkeys retain functional UOX(Oda, et al. 2002; Wu, et al. 1989; Wu, et al. 1992; Yeldandi,

et al. 1991). This case tests whether a PA matrix can distinguish a pseudogene retaining exon-level homology from a functional gene.

Here, we constructed an OrthoDB v12-derived PA matrix comprising 34 human genes involved in nitrogen waste excretion and ammonia metabolism across 60 species, totaling 2,040 gene-by-species cells. We applied an iterative audit framework integrating NCBI/RefSeq evidence, whole-genome tblastn, competitive paralog discrimination, locus-level clustering, phylogenetic priors, and published evidence. We evaluated whether cluster membership supports focal-ortholog presence, whether paralog misassignment can be distinguished from cluster bleed-through, and whether cell-level auditing can reveal lineage-specific gene gain, duplication, and pseudogenization.

2. Materials and Methods

2.1 Species Sampling and Gene Set Definition

We selected 60 species representing the major prokaryotic and eukaryotic lineages included in OrthoDB v12, with emphasis on taxa relevant to nitrogen-metabolism gene evolution and potentially anomalous orthology assignments. The panel comprised 6 prokaryotes, 7 unicellular eukaryotes or holozoans, 4 non-bilaterian metazoans, 8 invertebrate bilaterians, 17 non-mammalian vertebrates, and 18 mammals.

The prokaryotes included Bacillus subtilis, Escherichia coli, Mycobacterium tuberculosis, Methanocaldococcus jannaschii, Sulfolobus solfataricus, and Thermococcus kodakarensis. The archaeal species were included because of their membership in the GLUD2-associated OrthoDB cluster, allowing us to test whether deeply diverged glutamate dehydrogenase homologs could be misinterpreted as focal orthologs of the ape GLUD2 retrocopy. The panel therefore combined broad phylogenetic coverage with targeted sampling and was not randomly drawn from all species represented in OrthoDB.

The eukaryotic outgroups comprised Saccharomyces cerevisiae, Schizosaccharomyces pombe, Dictyostelium discoideum, Trypanosoma brucei, Monosiga brevicollis, Salpingoeca rosetta, and Capsaspora owczarzaki. Other taxa included one sponge, one ctenophore, two cnidarians, one echinoderm, one hemichordate, four protostomes, one cephalochordate, one tunicate, one cyclostome, one cartilaginous fish, four ray-finned fishes, three amphibians, four non-avian reptiles, and four birds.

The 18 mammals comprised one monotreme, two marsupials, and 15 placental mammals, including three rodents, five laurasiatherians, and seven primates. The primates included one prosimian, one New World monkey, one Old World monkey, and four great apes, including Homo sapiens. The complete species list and NCBI Taxonomy identifiers are provided in species_metadata.tsv.

The phylogenetic distribution of the sampled species is shown in Figure 1. Species-tree construction and divergence-time annotation are described in Section 2.12.

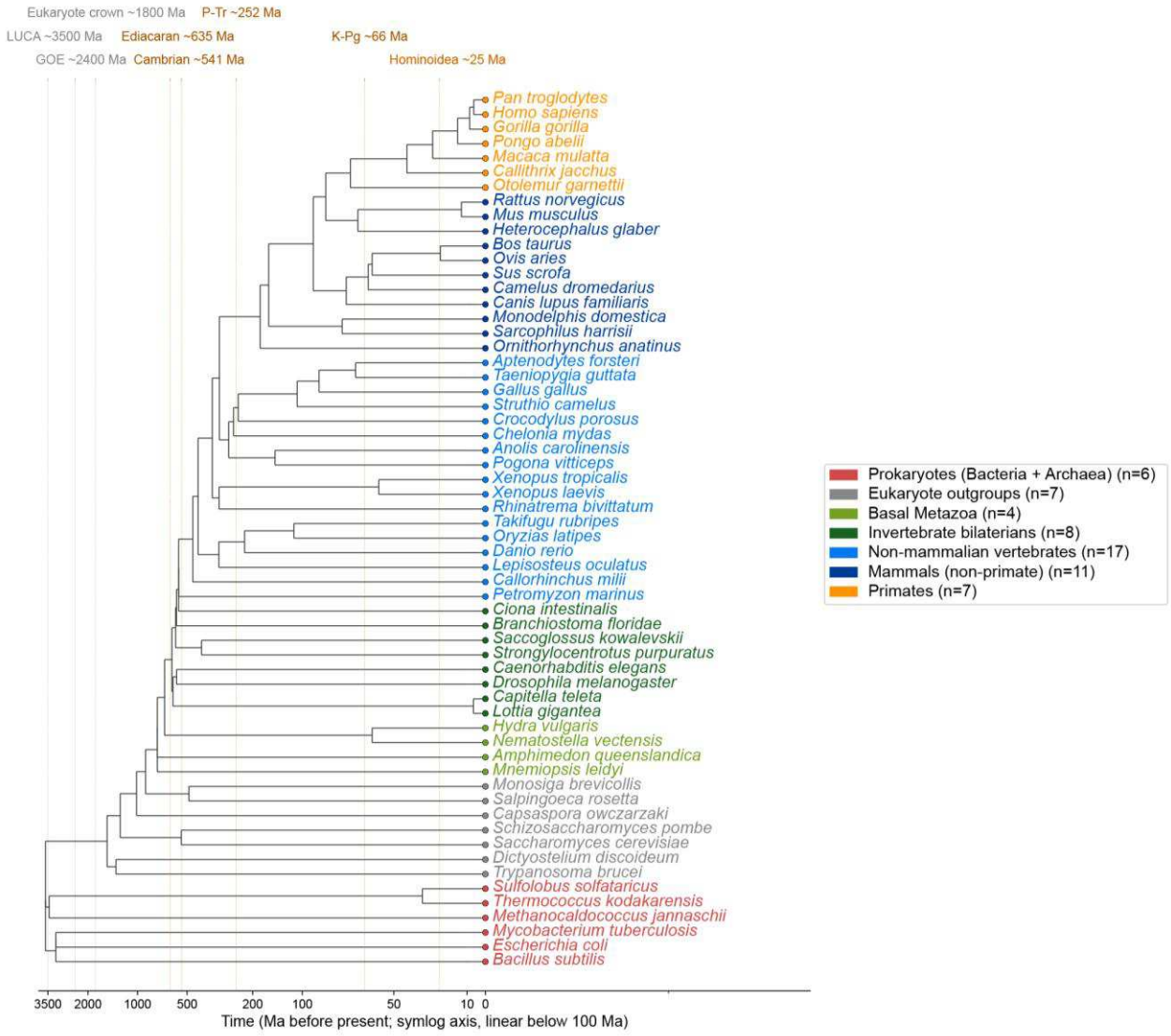


Figure 1. Time-scaled phylogenetic tree of the 60 sampled species. The topology was obtained from the Open Tree of Life, and divergence times were annotated using TimeTree v5. A symmetric logarithmic time axis was used. Geological boundaries and the Great Oxidation Event are visual temporal references only.

We selected 34 human genes involved in nitrogen-waste production, ammonia detoxification, glutamate metabolism, transamination, the urea cycle, urate metabolism, and transport of nitrogen-containing compounds: ABCG2, AQP1, AQP2, ARG1, ARG2, ASL, ASS1, CPS1, GLS, GLS2, GLUD1, GLUD2, GLUL, GOT1, GOT2, GPT, GPT2, NAGS, ODC1, OTC, PYCR1, RHAG,

RHBG, RHCG, SLC14A1, SLC14A2, SLC1A2, SLC1A3, SLC22A12, SLC25A13, SLC25A15, SLC2A9, UOX, and XDH.

The gene set included deeply conserved metabolic enzymes, transporters, lineage-specific paralogs, and pseudogenized genes. GLUD2 was included to evaluate orthology assignment for a recently derived retrocopy, whereas UOX was included to distinguish detectable sequence homology from functional-gene presence. Core machine-readable metadata for the 34 focal genes are provided in gene_metadata.tsv. Expanded functional and reference metadata are provided in network_proteins.tsv , and the UniProt/KEGG identifier verification results are provided in gene_metadata_verification_log.tsv.

2.2 Retrieval of OrthoDB Orthologous Groups and Initial PA Coding

We queried OrthoDB v12 for orthologous groups corresponding to the 34 human genes. A gene-by-species cell was initially coded as an OrthoDB-supported presence if the selected group contained at least one member from the corresponding species and as an OrthoDB-supported absence if no member was identified. This procedure produced an initial 34-gene × 60-species matrix containing 2,040 cells.

The initial calls were subsequently subjected to iterative cell-level auditing. Cells requiring review were identified from database inconsistencies, anomalous nomenclature, phylogenetically unexpected distributions, NCBI/RefSeq records, whole-genome tblastn results, competitive paralog discrimination, phylogenetic priors, and published evidence. Audited cells, supporting evidence, and final assignments are provided in pa_verification_log.tsv.

2.3 PA States, Evidence Types, and Display Codes

PA = 0.0: no support for the focal ortholog.

PA = 0.4: an ancestral homolog or paralog with unresolved orthology.

PA = 0.5: a known pseudogene.

PA = 1.0: presence supported solely by OrthoDB cluster membership.

PA = 1.1: presence additionally supported by NCBI/RefSeq records and sequence homology.

PA = 1.2: presence additionally supported by whole-genome tblastn.

PA = 1.3: OrthoDB cluster bleed-through.

The PA values are categorical labels rather than continuous probabilities or increasing levels of evidential strength. PA = 1.3 denotes an OrthoDB cluster member or gene-family-level homology signal that does not support focal-ortholog presence.

2.4 Phase 1: Cross-Checking NCBI/RefSeq Annotations

For cells showing discordance between OrthoDB and NCBI evidence, NCBI Entrez was queried using the target species and focal gene names, and retrieved records were processed using Biopython 1.84 (Cock, et al. 2009). Relevant RefSeq mRNA and protein records were compared with the human query and known close paralogs. Phase 1 was used to cross-check annotations rather than independently establish strict orthology.

Cells supported by both RefSeq records and sequence homology were assigned PA = 1.1 and recorded as refseq_gene_present. Cells annotated as pseudogenes or supported as inactive by published studies were assigned PA = 0.5. Unverifiable accessions were excluded as evidence for focal-gene presence and documented in the audit log.

The Callithrix jacchus GLUD2 mRNA accession XM_002753196, recorded in an earlier workflow, could not be verified and was excluded as evidence for GLUD2 presence.

The UOX status of the four great apes was assessed using NCBI sequences and published evidence. Cells retaining exon-level homology, carrying the shared codon 33 nonsense mutation, and

supported as inactive were assigned PA = 0.5 rather than being coded as functional-gene presences (Oda, et al. 2002; Wu, et al. 1989; Wu, et al. 1992; Yeldandi, et al. 1991).

2.5 Phase 2: Whole-Genome tblastn Searches

Human query proteins were searched against target-species genome assemblies for cells unresolved after Phase 1 and for all TIER 3 absence calls. Genome assemblies were obtained through NCBI Datasets (Sayers, et al. 2024), with assembly metadata provided in species_metadata.tsv. Local nucleotide databases were constructed using BLAST+ 2.16.0: `tblastn -query query_protein.fasta -db target_genome -evalue 1e-5 -outfmt 6 -max_target_seqs 50.`

Searches used BLOSUM62 and the default BLAST+ gap-opening and gap-extension penalties of 11 and 1, respectively (Altschul, et al. 1997; Camacho, et al. 2009; Henikoff and Henikoff 1992). Individual HSPs were required to meet E-value $\leq 1 \times 10^{-5}$, bit score ≥ 80, percent identity $\geq 30\%$, and aligned length ≥ 80 aa. At least one qualifying HSP was considered evidence of a homologous sequence signal, but not independent evidence of focal-ortholog identity.

Cells supported by the whole-genome search and not explained by a closer paralog were assigned PA = 1.2 and recorded as tblastn_present. Cells without qualifying HSPs were recorded as tblastn_absent. Sequences more consistent with ancestral homologs or other paralogs were assigned PA = 0.4 or PA = 0.0, as appropriate.

2.6 Locus-Level Clustering and Aggregate Scoring of tblastn HSPs

To avoid combining HSPs from distinct loci during chromosome-level aggregation, we applied single-linkage locus clustering (Sokal and Michener 1958). HSPs on the same subject sequence were sorted by subject start coordinate. Consecutive HSPs separated by no more than 100 kb were assigned to the same cluster; a new cluster was initiated when the distance exceeded 100 kb.

For each cluster, we calculated the summed HSP bit score, summed aligned length, and highest percent identity. The cluster with the highest sum_bitscore on each subject sequence was retained for subsequent scoring and paralog discrimination.

This procedure replaced an earlier chromosome-level method that could combine spatially separated loci. An SLC25A13 query against Schizosaccharomyces pombe produced 25 HSPs across an approximately 0.5 to 4.0 Mb interval. Chromosome-level aggregation yielded a cumulative aligned length of 3,276 aa for a 675-aa query, indicating the combination of distinct candidate loci.

For the 20-cell spot-check, locus clusters were evaluated using sum_bitscore ≥ 80, best_pident ≥ 30%, and sum_length ≥ 50 aa. Passing this criterion indicated a qualifying homologous sequence signal but did not independently establish focal-ortholog identity.

2.7 Competitive Paralog Discrimination

For gene families prone to cross-hits among close paralogs, the human focal protein and relevant close paralogs were independently used as tblastn queries at the same candidate locus. The families examined were GLUD1/GLUD2, SLC22A6/SLC22A12/SLC22A20, ARG1/ARG2, PYCR1/PYCR2/PYCR3, and SLC25A12/SLC25A13. SLC22A6, SLC22A20, PYCR2, PYCR3, and SLC25A12 were discrimination queries rather than focal genes.

If an alternative paralog produced a higher top-HSP bit score than the focal-gene query, the signal was attributed to the alternative paralog and not considered evidence for focal-ortholog presence. Because the procedure did not include a reverse search against the complete human proteome, it is termed competitive paralog discrimination rather than reciprocal best-hit analysis.

2.8 Phylogenetic Risk Stratification of Absence Calls

Calls assigned paralog_lineage_absence_orthodb were stratified according to the phylogenetic distance between the target species and the inferred origin of the focal gene or paralog. TIER 1

comprised 211 pre-bilaterian cells, TIER 2 comprised 81 invertebrate-bilaterian cells, and TIER 3 comprised four high-risk vertebrate cells. TIER 1 and TIER 2 represented relative levels of prior-based risk and were not treated as equivalent to direct sequence evidence.

The TIER 3 cells were GLUD2/Rhinatrema bivittatum, GLUD2/Sarcophilus harrisii, SLC22A12/R. bivittatum, and SLC22A12/Xenopus tropicalis. All four were evaluated through whole-genome tblastn and competitive paralog discrimination.

2.9 Analysis of GLUD1/GLUD2-Discriminating Sites

Human GLUD1 (UniProt P00367) and GLUD2 (UniProt P49448) were compared by Needleman-Wunsch global alignment using BLOSUM62 (Needleman and Wunsch 1970). Twenty-four positions containing different GLUD1 and GLUD2 residues were identified. For each position, the target residue was scored against the corresponding GLUD1 and GLUD2 residues, and the difference was calculated as $\Delta i = B(ai, rGLUD1,i) - B(ai, rGLUD2,i)$.

Under equal background frequencies for the 20 standard amino acids, the mean per-site standard deviation was 2.0533 raw BLOSUM62 units. Ten thousand empirical permutations using the same background and random seed 20260613 produced a per-site standard deviation of 2.0552, differing by 0.09%. The empirical and analytical standard deviations across all 24 sites were 4.134 and 4.152 bits, respectively.

Raw scores were converted to bits using $Sbits = \lambda Sraw/\ln 2$, where $\lambda = 0.267$ for gapped alignment with BLOSUM62 and the default BLAST+ gap penalties used here (Sheetlin, et al. 2005). Because the analytical and empirical calculations used the same uniform background, their agreement validates the implementation rather than the background assumption.

For a candidate locus covering k discriminating positions, the expected noise scale was $\sigma k = \sigma site\sqrt{k}$, and the standardized score was $Z = (\Delta observed - \mu k)/\sigma k$. Positive values favored GLUD1

and negative values favored GLUD2. Values exceeding 2σ were treated as strong discrimination signals; lower values in a consistent direction were treated as directional support.

2.10 Random Spot-Check of OrthoDB-Only Presence Calls

Twenty cells were selected by simple random sampling without replacement from the 1,580 PA = 1.0 cells present at the time of sampling, using random seed 20260609. Sampling preceded the reclassification of GLUD2/Sulfolobus solfataricus to PA = 1.3 and Gorilla gorilla UOX to PA = 0.5. These revisions reduced the final PA = 1.0 total to 1,578 but did not affect the sampled cells.

Each cell underwent whole-genome tblastn and locus-level HSP clustering. The strict single-HSP criterion required bit score ≥ 80, percent identity ≥ 30%, and aligned length ≥ 80 aa. The locus-aggregate criterion required sum_bitscore ≥ 80, best_pident ≥ 30%, and sum_length ≥ 50 aa. The criteria differed for GOT2/Oryzias latipes and ABCG2/Xenopus laevis, in which high-identity signals were distributed across multiple HSPs. The spot-check assessed sequence-level recoverability rather than strict 1:1 orthology.

2.11 Statistical Analysis

Binomial proportions were reported with 95% Wilson confidence intervals without continuity correction (Wilson 1927). Wilson intervals were used because several analyses involved small samples. Calculations were performed using SciPy 1.13 (Virtanen, et al. 2020) and statsmodels 0.14.

2.12 Species-Tree Construction and Divergence-Time Annotation

The topology of the 60-species tree was obtained from the Open Tree of Life. Divergence times were retrieved from TimeTree v5 through the TimeTree REST API(Kumar, et al. 2017). A symmetric logarithmic time axis was used to display both recent and deep events. Geological

boundaries and the Great Oxidation Event were visual references and were not used to infer topology or node ages.

2.13 Reconstruction of the GLUD1/GLUD2 Gene Tree

The GLUD1/GLUD2 gene tree comprised 51 protein sequences, including 45 GLUD1 or GLUD1-like sequences and 6 GLUD2 or GLUD2-related sequences. The sequence set represented major vertebrate lineages and included glutamate dehydrogenase sequences from Drosophila and Ciona as deep gene-family references. The maximum-likelihood tree was constructed using IQ-TREE 2.

2.14 Extraction and Alignment of the UOX Codon 33 Region

Sequences surrounding codon 33 in UOX exon 2 were retrieved from M69165 (Pan troglodytes), M69167 (Pongo sp.), M69168 (Macaca mulatta), AH005326 (Gorilla gorilla), and NR_003927.2 (Homo sapiens). For each taxon, a 27-bp region comprising 12 bp upstream, the focal codon, and 12 bp downstream was extracted. Macaca mulatta, which retains functional UOX, was used as the outgroup reference. M69167 was retained as Pongo sp., following its original GenBank annotation.

2.15 Data and Code Availability

The principal data files include species_metadata.tsv, gene_metadata.tsv, presence_absence_matrix_verified.tsv, pa_verification_log.tsv, pa1_spotcheck_aggregate_v2.tsv, glud1_glud2_discriminating_sites.tsv, glud_disc_site_permutation_null.tsv, glud_permutation_null_distribution_10k.tsv, glud_locus_bitscore_delta_noise_analysis_v3.tsv, and uox_codon33_alignment.fasta.

3. Results

3.1 The Audited 34-Gene × 60-Species PA Matrix

The final matrix contained 412 cells without support for the focal ortholog, 6 ancestral or unresolved homologs, 5 known pseudogenes, 1,578 OrthoDB-only presence calls, 30 NCBI/RefSeq-supported presence calls, 8 tblastn-supported presence calls, and 1 cluster bleed-through call, totaling 2,040 cells. PA = 1.0 to PA = 1.2 represent distinct evidence sources rather than increasing probabilities or evidential strength. PA = 1.3 represents cluster bleed-through. Seventeen of the 34 genes contained at least one nonbinary PA state. The complete PA matrix is shown in Figure 2.

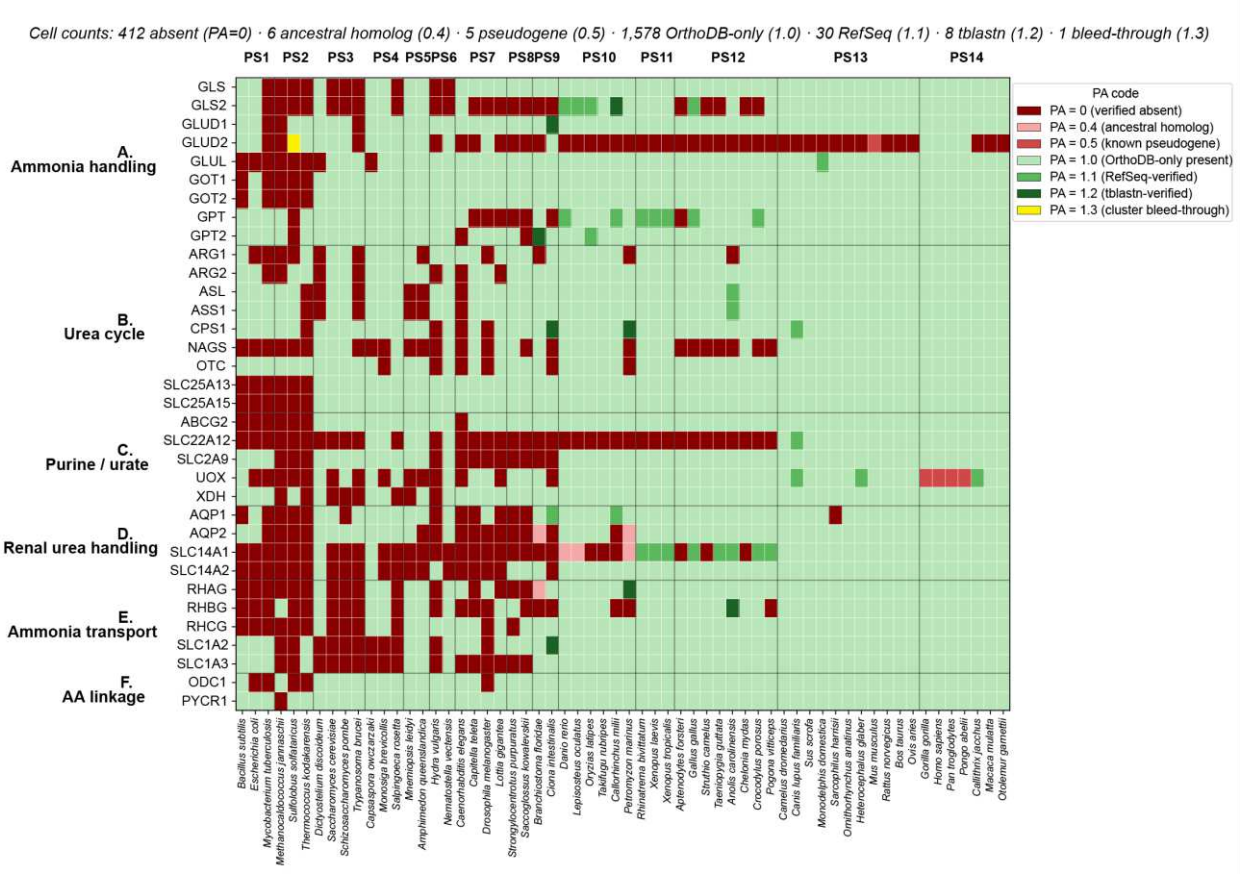


Figure 2. Audited 34-gene × 60-species presence/absence matrix. Rows represent focal genes and columns represent species. PA = 0.0 indicates no support for the focal ortholog; PA = 0.4,

ancestral or unresolved homology; PA = 0.5, a known pseudogene; PA = 1.0, OrthoDB-only presence; PA = 1.1, NCBI/RefSeq-supported presence; PA = 1.2, tblastn-supported presence; and PA = 1.3, cluster bleed-through.

The matrix retained broad phylogenetic patterns. CPS1, ASS1, and ASL were present in all 35 vertebrates examined. OTC was present in 34/35 vertebrates, whereas ARG1 was present in 33/35. GLUD2 was supported only in the four great apes and not in the 14 non-ape mammals or 17 non-mammalian vertebrates. Signals from prokaryotic and deeply diverged eukaryotic GLUD2-cluster members were more consistent with ancient GDH-family homologs or GLUD1-like paralogs.

SLC22A12 was present in all 18 mammals but lacked focal-ortholog support in the 17 non-mammalian vertebrates. GPT2 and GPT were coded as present in 57/60 and 52/60 species, respectively.

UOX was coded as PA = 0.5 in Pongo abelii, Gorilla gorilla, Pan troglodytes, and Homo sapiens. Macaca mulatta, Callithrix jacchus, and Otolemur garnettii retained UOX calls consistent with functional genes, in agreement with pseudogenization in the Hominidae ancestor.

Gene-family age analysis assigned 24 genes to families with deep prokaryotic homologs, 9 to families recovered primarily in eukaryotes, and SLC22A12 to a Holozoa-associated family. GLUD2 was treated separately because the GDH family is ancient, whereas the GLUD2 retrocopy originated approximately 18 to 25 Ma (Figure 3).

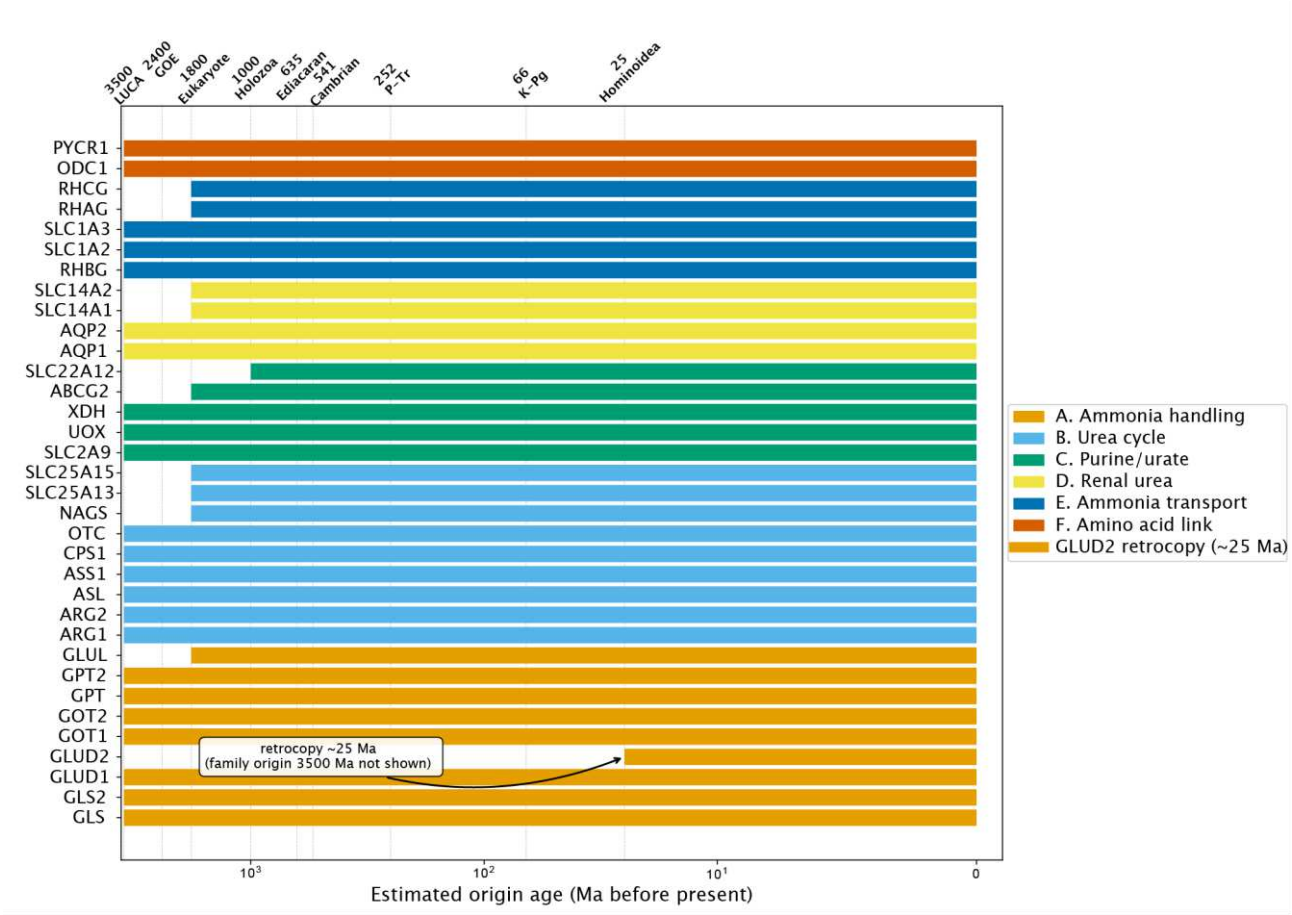


Figure 3. Phylostratigraphic distribution of 34 nitrogen-metabolism-related genes. Bars show estimated gene-family origin times. The GLUD2 bar represents the origin of the ape retrocopy rather than the deeper origin of the GDH family.

3.2 Iterative Cell-Level Auditing and Evidence Composition

The audit expanded as annotation conflicts, phylogenetically unexpected calls, and additional sequence evidence were identified. In total, 462 unique cells underwent at least one cell-level review, whereas 1,578 retained OrthoDB-only presence calls.

The largest evidence category comprised 296 paralog_lineage_absence_orthodb cells in which cluster membership was inconsistent with the inferred origin of the focal gene and the phylogenetic position of the target species. Other audited cells were classified according to RefSeq support,

whole-genome tblastn results, paralog discrimination, pseudogene evidence, or cluster bleed-through (Figure 4).

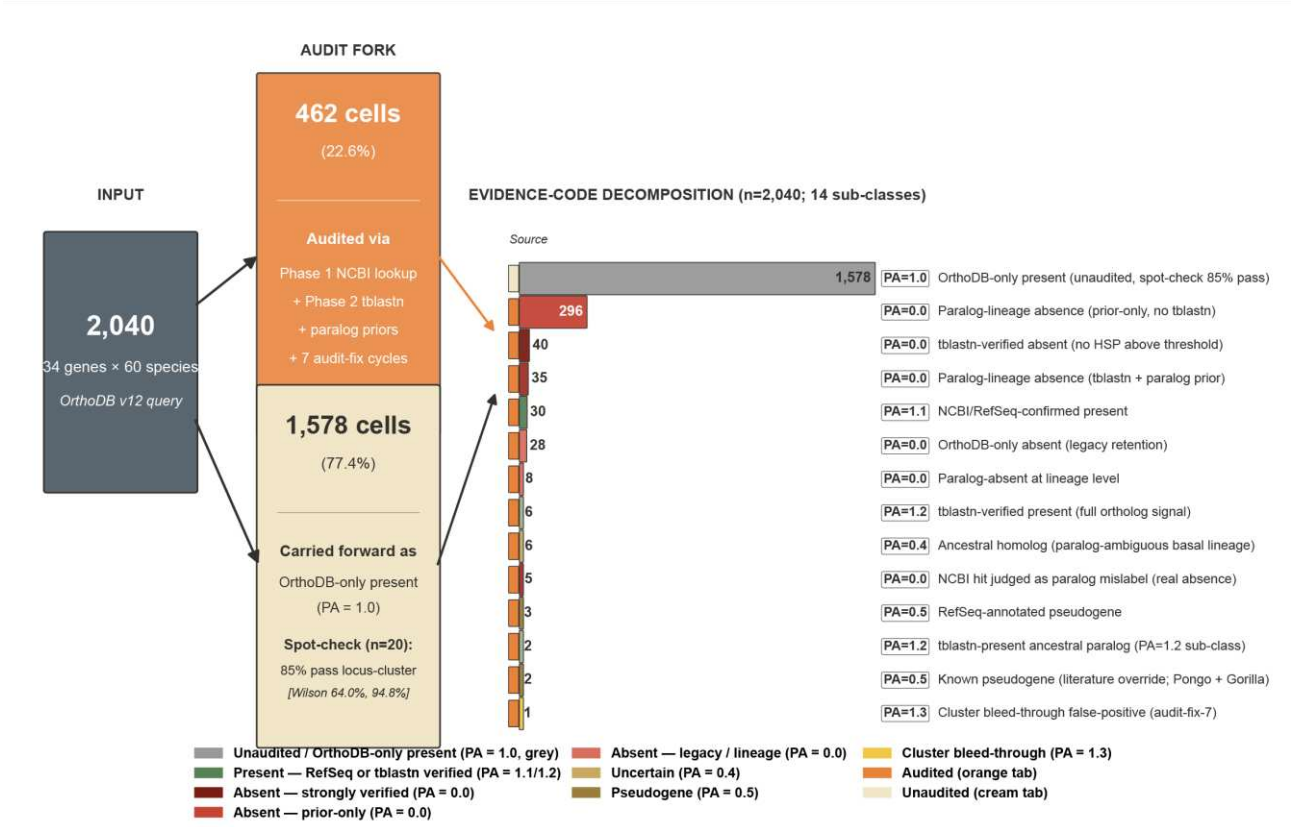


Figure 4. Cell-level audit coverage and final evidence composition. Of the 2,040 cells, 462 underwent cell-level review and 1,578 retained OrthoDB-only presence calls. Green indicates evidence supporting focal-ortholog presence, red absence, amber ancestral or unresolved homology, yellow cluster bleed-through, and gray unaudited OrthoDB-only presence.

The audit removed or revised accessions and gene names that could not be supported by the records examined, reassigned paralog-derived calls in nontarget lineages, replaced chromosome-level aggregation with locus-level clustering, and directly evaluated high-risk absence calls and randomly sampled OrthoDB-only presence calls.

3.3 Direct Validation of High-Risk Absence Calls

The 296 prior-supported absence cells were divided into 211 TIER 1, 81 TIER 2, and 4 TIER 3 cells. The TIER 3 cells were GLUD2/Rhinatrema bivittatum, GLUD2/Sarcophilus harrisii, SLC22A12/R. bivittatum, and SLC22A12/Xenopus tropicalis.

The SLC22A12-associated signals in both amphibians were better explained by SLC22A6/A20-like paralogs than by SLC22A12. Candidate GLUD loci in R. bivittatum and S. harrisii favored GLUD1 over GLUD2 by 3.92σ and 5.58σ, respectively (Figure 5, Figure S1). Direct sequence evidence therefore supported all four highest-risk absence assignments, corresponding to 4/4 with a Wilson 95% confidence interval of 51.0% to 100.0%.

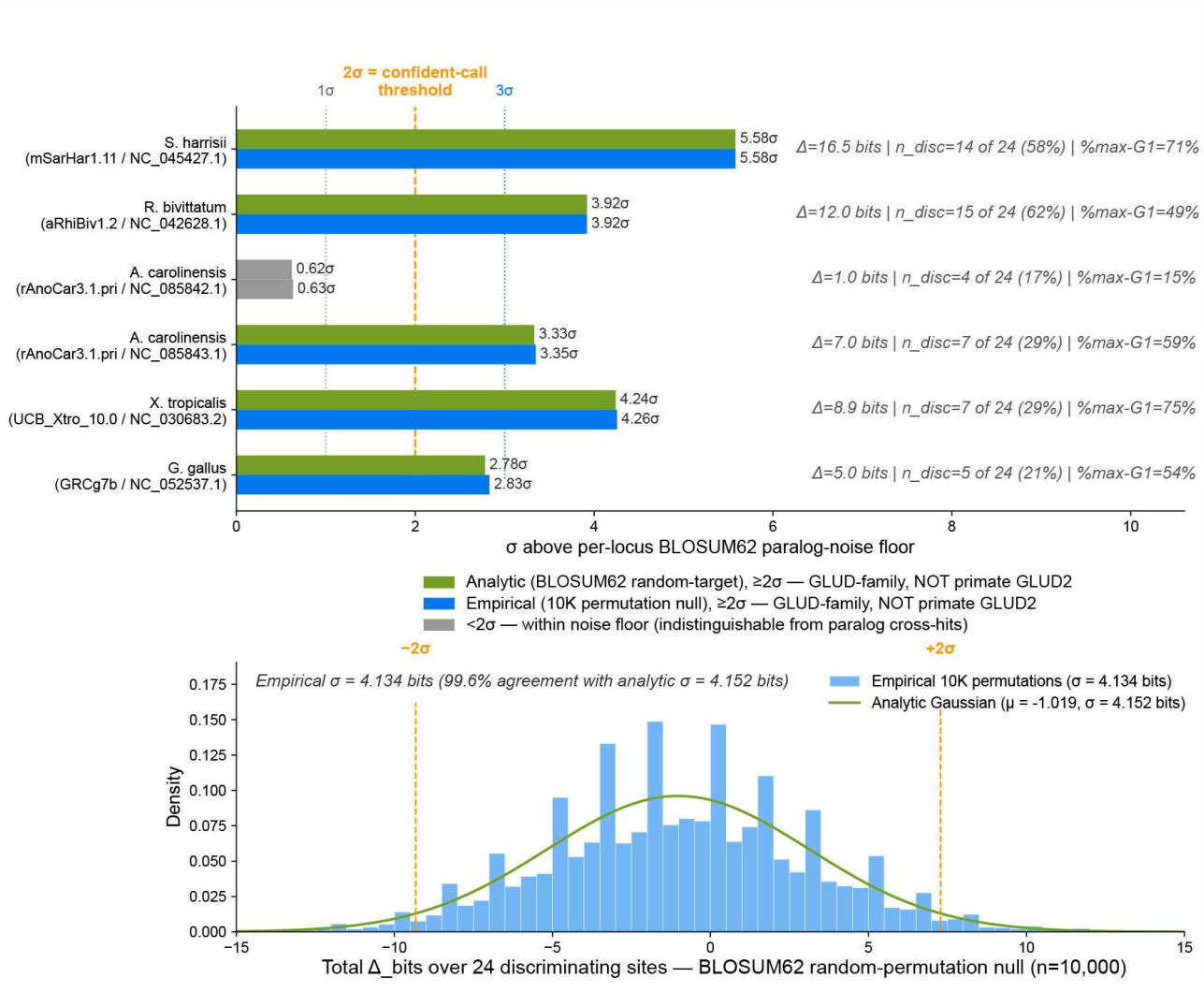


Figure 5. GLUD1/GLUD2 discrimination at six candidate GLUD loci from non-ape vertebrates. Positive standardized scores favor GLUD1 and negative scores favor GLUD2. The 2σ threshold indicates a strong signal. Analytical and empirical scales are shown, together with the corresponding null distributions.

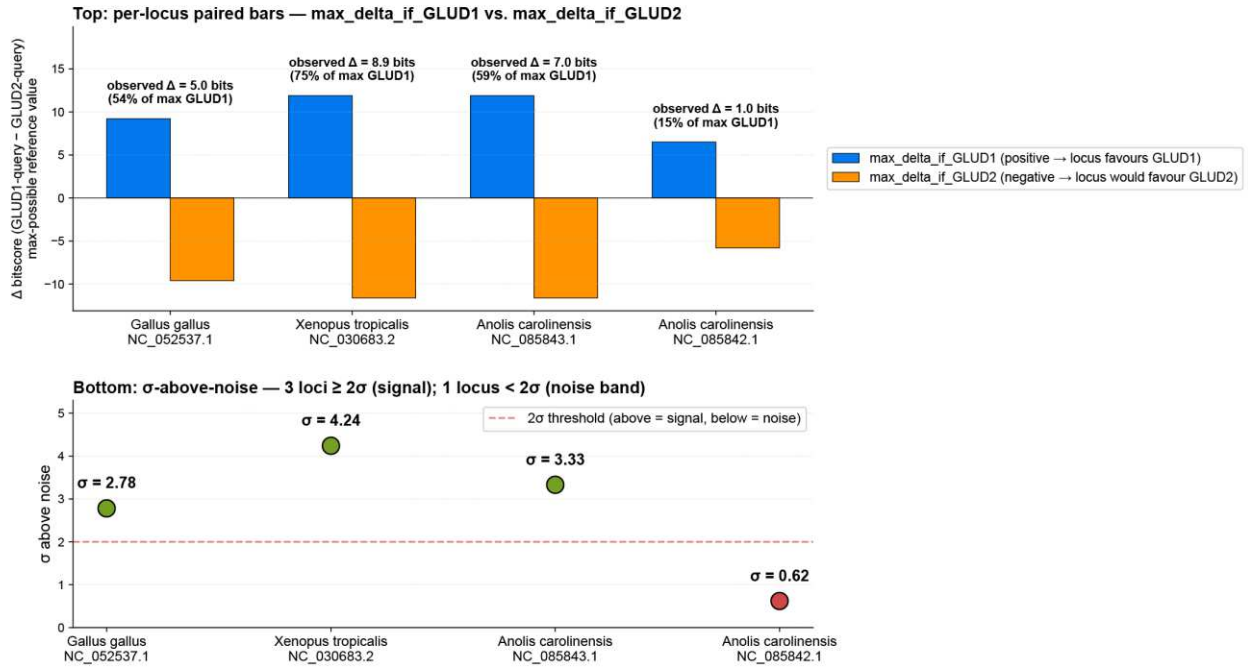


Figure S1. Four non-ape vertebrate GLUD candidate loci included in the initial analysis. The loci were sampled from Gallus gallus, Xenopus tropicalis, and Anolis carolinensis. Three exceeded the 2σ discrimination threshold, whereas A. carolinensis NC_085842.1 favored GLUD1 directionally but remained below the threshold.

3.4 Lineage-Specific Evolutionary Events Revealed by the Audit

3.4.1 Non-Ape Candidate GLUD Loci Favor GLUD1

Six candidate loci from five non-ape vertebrate species were examined. All six favored GLUD1 over GLUD2, corresponding to 6/6 with a Wilson 95% confidence interval of 61.0% to 100.0%.

Five exceeded the 2σ threshold, whereas none favored GLUD2. The GLUD1/GLUD2 maximum-likelihood gene tree placed the GLUD2 sequences from four great apes within the vertebrate GLUD1 lineage(Figure 6). Together, these results support the interpretation that non-ape GLUD2-associated cluster members represent GLUD1, GLUD1-like sequences, or deeper GDH-family homologs rather than orthologs of the ape GLUD2 retrocopy.

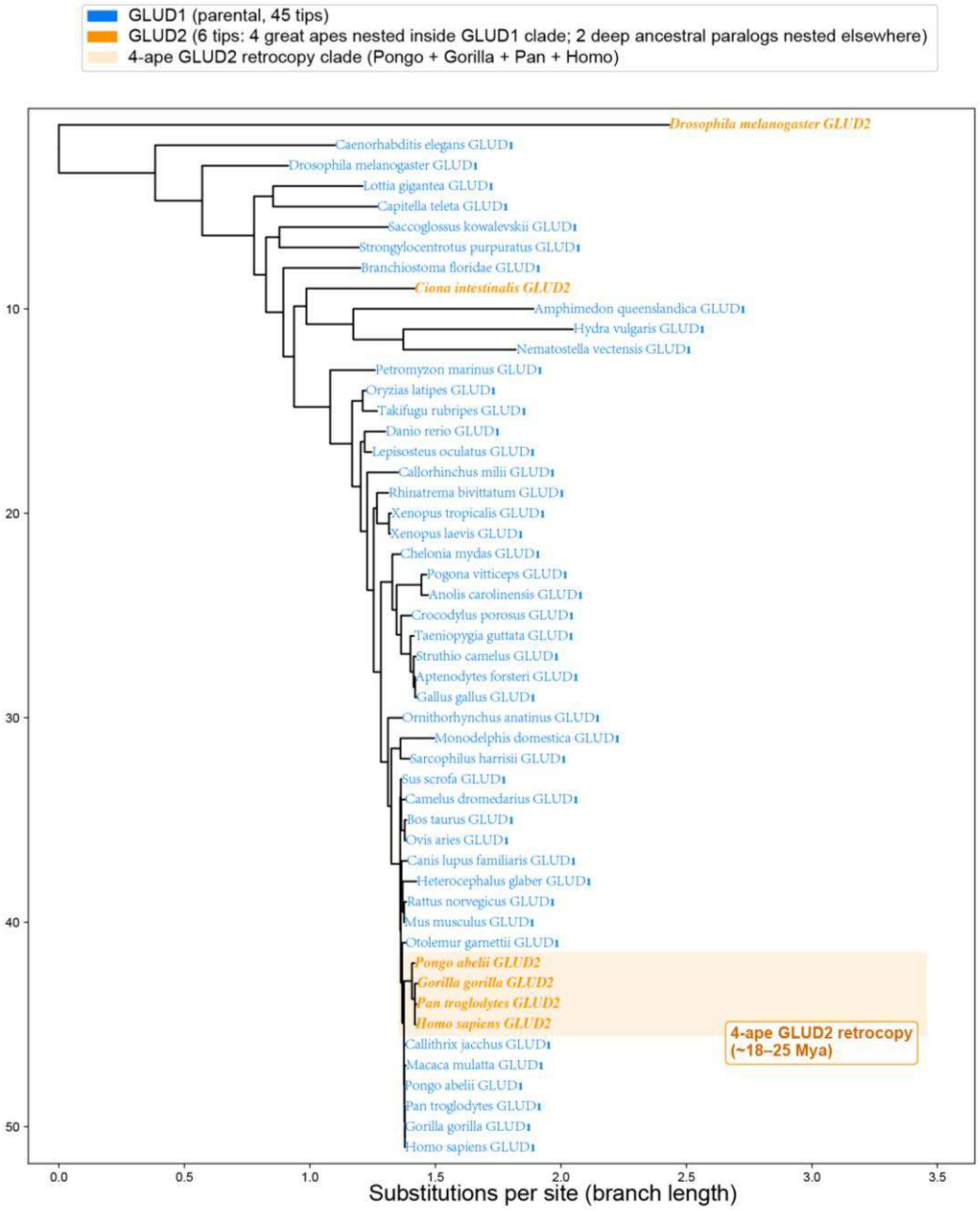


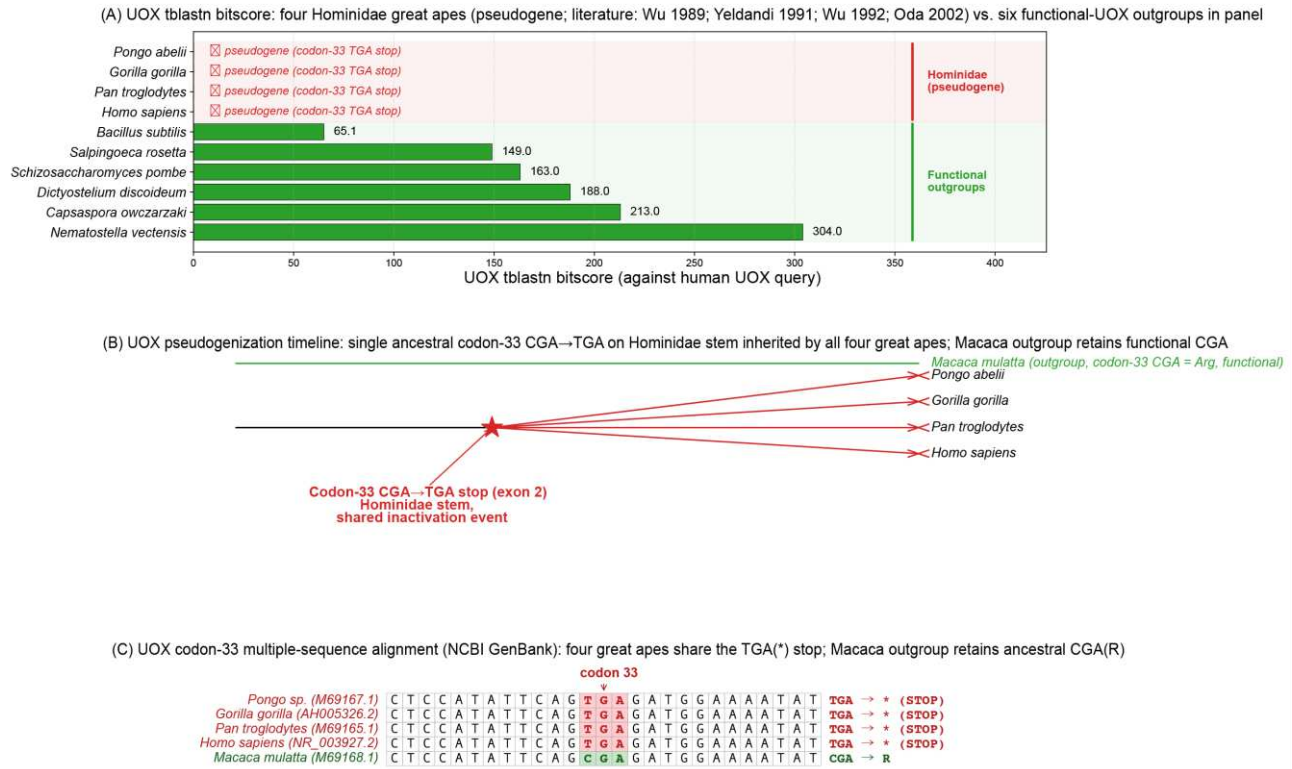
Figure 6. Maximum-likelihood gene-family tree of GLUD1 and GLUD2. The tree includes 45 GLUD1 or GLUD1-like sequences and 6 GLUD2 or GLUD2-related sequences. Great-ape GLUD2 sequences form a clade within the vertebrate GLUD1 lineage.

3.4.2 Candidate GLUD1 Duplication in Anolis carolinensis

Two GLUD1-like loci were recovered on NC_085842.1 and NC_085843.1 in Anolis carolinensis. NC_085843.1 covered 7/24 discriminating positions and favored GLUD1 by 3.33σ. NC_085842.1 covered 4/24 positions and favored GLUD1 directionally but did not exceed 2σ. These loci suggest a candidate GLUD1 duplication in the Anolis lineage. Because only one lepidosaur species was directly examined and one locus had limited coverage, the event was not assigned to Lepidosauria as a whole.

3.4.3 Shared UOX Codon 33 Stop Mutation in Great Apes

The four great apes retained exon-level UOX homology and shared a TGA stop codon at codon 33, whereas Macaca mulatta retained the ancestral CGA arginine codon (Figure 7). This pattern supports UOX pseudogenization in the common ancestor of Hominidae. The four great-ape cells were coded as PA = 0.5 to distinguish retained pseudogene sequences from functional genes and complete genomic absence.



365

366 **Figure 7. Great-ape UOX pseudogenization and the codon 33 nonsense mutation.** Four great
367 apes retain exon-level UOX homology and share the codon 33 CGA-to-TGA substitution, whereas
368 Macaca mulatta retains the ancestral CGA codon.

369 **3.5 Random Spot-Check of OrthoDB-Only Presence Calls**

370 All 20 sampled cells contained at least one HSP with an E-value $\leq 1 \times 10^{-5}$. Fifteen passed the strict
371 single-HSP criterion, corresponding to 75% with a Wilson 95% confidence interval of 53.1% to
372 88.8%. Seventeen passed the locus-aggregate criterion, corresponding to 85% with a Wilson 95%
373 confidence interval of 64.0% to 94.8%.

374 The criteria differed for GOT2/Oryzias latipes and ABCG2/Xenopus laevis, both of which
375 contained high-identity signals distributed across multiple HSPs. Three cells did not pass the locus-

aggregate criterion: SLC1A3/Mycobacterium tuberculosis, OTC/Trypanosoma brucei, and SLC25A13/Salpingoeca rosetta.

GLUD2/Sulfolobus solfataricus passed the locus-aggregate criterion, but its sequence signal represented an archaeal GDH-family homolog rather than the ape GLUD2 retrocopy. It was therefore classified as cluster bleed-through(Figure 8). The spot-check assessed sequence-level recoverability rather than strict 1:1 orthology.

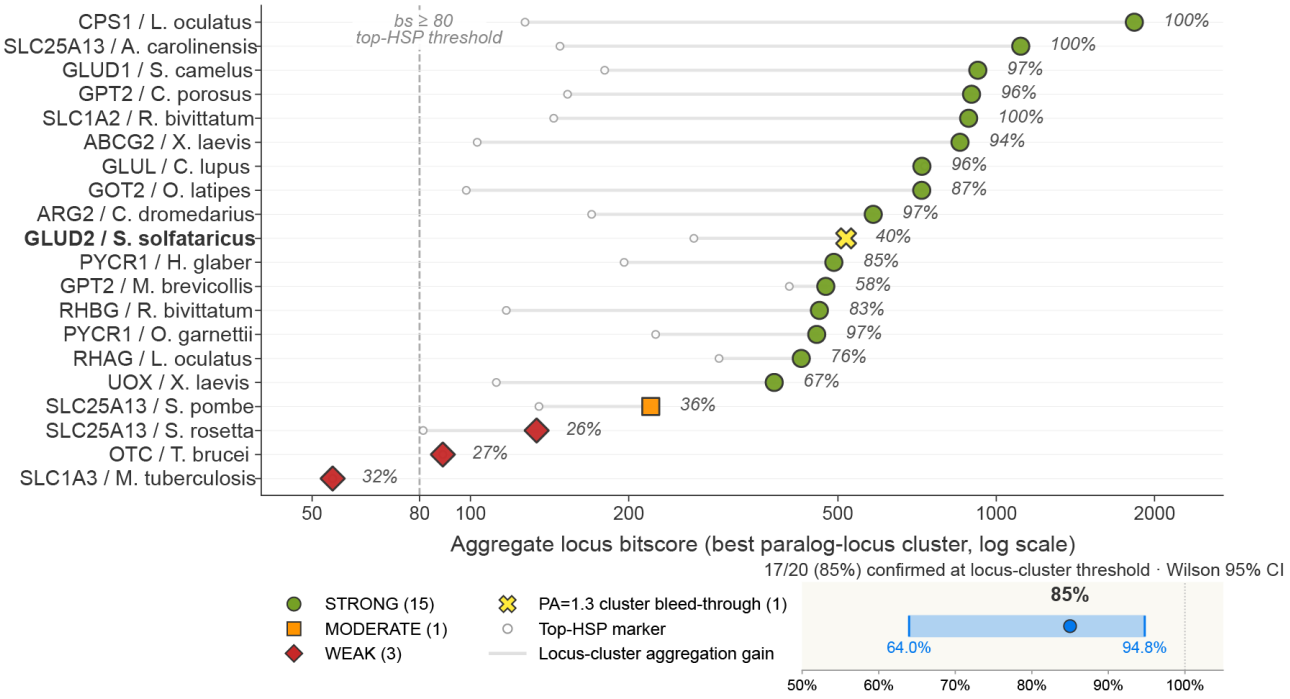


Figure 8. Whole-genome sequence assessment of 20 randomly sampled OrthoDB-only presence cells. The figure shows results under the strict single-HSP and locus-aggregate criteria. GLUD2/Sulfolobus solfataricus is separately classified as PA = 1.3 cluster bleed-through.

3.6 Cluster Bleed-Through and Paralog Misassignment

GLUD2/Sulfolobus solfataricus produced a best locus-cluster bit score of 518.0, a best_pident of 39.6%, and a highest single-HSP bit score of 266. Because S. solfataricus is an archaeon and

GLUD2 arose in the ape ancestor, the signal cannot represent an ortholog of the ape GLUD2 retrocopy.

Cluster bleed-through occurs when a deeply diverged homolog or ancient paralog is placed in the same database cluster as a recently derived focal gene. Paralog misassignment instead occurs when a different close paralog in the target species is interpreted as the focal ortholog. GLUD2/S. solfataricus represents cluster bleed-through, whereas SLC22A6/A20-like signals recovered with an SLC22A12 query represent paralog misassignment. The GLUD2/S. solfataricus cell was assigned PA = 1.3 and recorded as orthodb_cluster_bleed_through.

4. Discussion

4.1 Evidence Heterogeneity in Database-Derived PA Matrices

The audit demonstrates that database cluster membership, detectable sequence homology, focal-ortholog identity, and functional-gene presence represent related but nonequivalent levels of inference. Collapsing these levels into binary PA states may distort estimates of gene origin, phylogenetic distribution, and functional capacity. The categorical PA scheme retains distinctions among the main biological interpretations and evidence types used in the audit.

Of the 2,040 cells, 462 underwent cell-level review, whereas 1,578 retained OrthoDB-only presence calls. The matrix should therefore be interpreted as evidence-stratified rather than uniformly validated. Likewise, the 75% and 85% confirmation rates depend on the single-HSP and locus-aggregate criteria. They reflect sequence-level recoverability under the specified criteria, not strict 1:1 orthology or overall OrthoDB accuracy.

4.2 Cluster Bleed-Through and Paralog Misassignment

Paralog misassignment and cluster bleed-through represent distinct sources of apparent PA false positives. Paralog misassignment occurs when another close paralog in the target genome produces a stronger signal than the focal gene(Dessimoz, et al. 2012; Sonnhammer, et al. 2014). The SLC22A12-associated signals attributed to SLC22A6, SLC22A20, or related transporters exemplify this mechanism.

Cluster bleed-through arises when deeply diverged homologs or ancient paralogs are included in the same database cluster as a recently derived focal gene. GLUD2/S. solfataricus exemplifies this case because its signal reflects an archaeal GDH-family homolog rather than the ape GLUD2 retrocopy, which originated from GLUD1 in the ape ancestor (Burki and Kaessmann 2004; Rosso, et al. 2008). PA = 1.3 preserves the original evidence while indicating that the signal does not support focal-ortholog identity.

4.3 Evolutionary Implications

All six non-ape GLUD candidate loci favored GLUD1 over GLUD2, and five exceeded the 2σ threshold. Together with the placement of great-ape GLUD2 sequences within the vertebrate GLUD1 lineage, these findings support the ape-specific retrocopy origin of GLUD2. The limited number of species and the shared uniform amino-acid background assumption nevertheless warrant cautious interpretation.

Two GLUD1-like loci in Anolis carolinensis suggest a candidate GLUD1 duplication in the Anolis lineage. Broader lepidosaur sampling, open-reading-frame assessment, synteny analysis, and gene-tree placement are required before the duplication can be assigned to a specific ancestral branch.

Great-ape UOX illustrates a different limitation of binary PA coding. The four Hominidae species retained exon-level homology and shared the codon 33 nonsense mutation, supporting

pseudogenization rather than functional-gene presence or complete genomic absence. Together, GLUD2, the candidate Anolis GLUD1 duplication, and UOX show how gene gain, duplication, and pseudogenization can generate deep-homolog co-clustering, within-cluster paralog ambiguity, and sequence retention despite functional loss.

4.4 Limitations

This study began with pre-existing OrthoDB groups and did not reconstruct all gene families de novo. Orthologs omitted from the selected clusters may therefore have remained undetected. The results of tblastn and competitive paralog discrimination may also be affected by assembly quality, sequence divergence, gene-model completeness, and threshold selection.

The 20-cell spot-check was small and unstratified and cannot provide a precise estimate of the error rate across the complete matrix. The PA matrix also reflects genomic potential rather than direct evidence of gene expression, protein function, or metabolic phenotype.

4.5 Generalizability and Future Directions

Future studies could combine larger stratified samples with candidate-protein reconstruction, reverse proteome searches, gene-tree analysis, and synteny assessment to distinguish focal orthologs, paralogs, and pseudogenes more rigorously. The resulting matrix could support downstream analyses of selection pressure, gene gain and loss, and correlated trait evolution.

The framework is applicable to other database-derived PA matrices containing lineage-specific gene gains, duplications, pseudogenization events, or losses. Compared with treating database output as binary ground truth, cell-level auditing provides a more transparent and traceable basis for comparative genomic analysis.

5. Conclusion

We constructed and audited a 34-gene × 60-species PA matrix comprising 2,040 gene-by-species cells. The final matrix contained 412 cells without support for the focal ortholog, 6 ancestral or unresolved homologs, 5 known pseudogenes, 1,578 OrthoDB-only presence calls, 38 presence calls additionally supported by sequence or annotation evidence, and 1 cluster bleed-through call. The 20-cell spot-check yielded confirmation rates of 75% and 85% under the strict single-HSP and locus-aggregate criteria, respectively, and direct sequence evidence supported all four highest-risk absence assignments.

The analysis supported the ape-specific retrocopy origin of GLUD2, identified a candidate GLUD1 duplication in Anolis carolinensis, and confirmed the shared UOX codon 33 pseudogenization marker in four great apes. Database cluster membership, gene-family homology, focal-ortholog identity, and functional-gene presence are therefore not equivalent. Cell-level auditing provides a more transparent and traceable basis for interpreting orthology-database-derived PA matrices.

Funding

This work was supported by the First-Class Academic Subjects Special Research Project of the Education Department of Inner Mongolia Autonomous Region (YLXKZX-NND-031); the Capacity Building Project of the Grassland Rodent Ecology and Rodent Pest Control Key Laboratory of Higher Education Institutions of the Inner Mongolia Autonomous Region (RZ2500000852); the Double First-Class Construction Funds, Discipline-Specific Research Project in Grass Science (DC2500000964); the National Key Research and Development Program of China (2024YFD1400503); the Key Research and Development Program of Inner Mongolia Autonomous Region (2025YFHH0177); the Key Program of the Natural Science Foundation of Inner Mongolia Autonomous Region (2025SHZR0901); the Major Program of the National Natural Science

Foundation of China (32090024); the Fundamental Research Funds for Universities Directly Affiliated with the Inner Mongolia Autonomous Region (BR221307, BR220106, and BR221037); and the Special Project for First-Class Discipline Research (YLXKZX-NND-005).

Author Contributions

Y.L. and H.F. conceived the study. Y.L. performed the computational analyses and prepared the original manuscript. Both authors reviewed and revised the manuscript.

Ethics declarations

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Data Availability

The audited PA matrix, cell-level audit log, spot-check results, GLUD1/GLUD2 discrimination files, UOX codon 33 alignment, and analysis scripts will be deposited in a permanent public repository. Public genome assemblies and sequence records analyzed in this study are identified in the accompanying metadata files.

Figure Legends

Figure 1. Time-scaled phylogenetic tree of the 60 sampled species. The topology was obtained from the Open Tree of Life, and divergence times were annotated using TimeTree v5. A symmetric logarithmic time axis was used. Geological boundaries and the Great Oxidation Event are visual temporal references only.

Figure 2. Audited 34-gene × 60-species presence/absence matrix. Rows represent focal genes and columns represent species. PA = 0.0 indicates no support for the focal ortholog; PA = 0.4, ancestral or unresolved homology; PA = 0.5, a known pseudogene; PA = 1.0, OrthoDB-only presence; PA = 1.1, NCBI/RefSeq-supported presence; PA = 1.2, tblastn-supported presence; and PA = 1.3, cluster bleed-through.

Figure 3. Phylostratigraphic distribution of 34 nitrogen-metabolism-related genes. Bars show estimated gene-family origin times. The GLUD2 bar represents the origin of the ape retrocopy rather than the deeper origin of the GDH family.

Figure 4. Cell-level audit coverage and final evidence composition. Of the 2,040 cells, 462 underwent cell-level review and 1,578 retained OrthoDB-only presence calls. Green indicates evidence supporting focal-ortholog presence, red absence, amber ancestral or unresolved homology, yellow cluster bleed-through, and gray unaudited OrthoDB-only presence.

Figure 5. GLUD1/GLUD2 discrimination at six candidate GLUD loci from non-ape vertebrates. Positive standardized scores favor GLUD1 and negative scores favor GLUD2. The 2σ threshold indicates a strong signal. Analytical and empirical scales are shown, together with the corresponding null distributions.

Figure 6. Maximum-likelihood gene-family tree of GLUD1 and GLUD2. The tree includes 45 GLUD1 or GLUD1-like sequences and 6 GLUD2 or GLUD2-related sequences. Great-ape GLUD2 sequences form a clade within the vertebrate GLUD1 lineage.

Figure 7. Great-ape UOX pseudogenization and the codon 33 nonsense mutation. Four great apes retain exon-level UOX homology and share the codon 33 CGA-to-TGA substitution, whereas Macaca mulatta retains the ancestral CGA codon.

Figure 8. Whole-genome sequence assessment of 20 randomly sampled OrthoDB-only presence cells. The figure shows results under the strict single-HSP and locus-aggregate criteria. GLUD2/Sulfolobus solfataricus is separately classified as PA = 1.3 cluster bleed-through.

Figure S1. Four non-ape vertebrate GLUD candidate loci included in the initial analysis. The loci were sampled from Gallus gallus, Xenopus tropicalis, and Anolis carolinensis. Three exceeded the 2σ discrimination threshold, whereas A. carolinensis NC_085842.1 favored GLUD1 directionally but remained below the threshold.

References

Altenhoff, A. M., et al.

2024 OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. Nucleic Acids Res 52(D1):D513-d521.

Altschul, S. F., et al.

1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389-402.

Burki, F., and H. Kaessmann

2004 Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet 36(10):1061-3.

Camacho, C., et al.

2009 BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Cock, P. J., et al.

2009 Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11):1422-3.

Dessimoz, C., et al.

2012 Toward community standards in the quest for orthologs. Bioinformatics 28(6):900-4.

Dimovasili, C., et al.

2021 Crystal structure of glutamate dehydrogenase 2, a positively selected novel human enzyme involved in brain biology and cancer pathophysiology. J Neurochem 157(3):802-815.

Henikoff, S., and J. G. Henikoff

1992 Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22):10915-9.

Hernandez-Plaza, A., et al.

2023 eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res 51(D1):D389-d394.

Kumar, S., et al.

2017 TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34(7):1812-1819.

Kuznetsov, D., et al.

2023 OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res 51(D1):D445-d451.

Litso, I., et al.

2023 Structural Evolution of Primate Glutamate Dehydrogenase 2 as Revealed by In Silico Predictions and Experimentally Determined Structures. Biomolecules 14(1).

Needleman, S. B., and C. D. Wunsch

1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443-53.

Oda, M., et al.

2002 Loss of urate oxidase activity in hominoids and its evolutionary implications. Mol Biol Evol 19(5):640-53.

Plaitakis, A., et al.

2017 The Glutamate Dehydrogenase Pathway and Its Roles in Cell and Tissue Biology in Health and Disease. Biology (Basel) 6(1).

Rosso, L., et al.

2008 Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive Darwinian selection. PLoS Genet 4(8):e1000150.

Sayers, E. W., et al.

2024 Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 52(D1):D33-d43.

Sheetlin, S., Y. Park, and J. L. Spouge

2005 The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment. Nucleic Acids Res 33(15):4987-94.

Sokal, Robert R, and Charles D Michener

1958 A statistical method for evaluating systematic relationships.

Sonnhammer, E. L., et al.

2014 Big data and other challenges in the quest for orthologs. Bioinformatics 30(21):2993-8.

Virtanen, P., et al.

2020 SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17(3):261-272.

Wilson, Edwin B

1927 Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22(158):209-212.

Wu, X. W., et al.

1989 Urate oxidase: primary structure and evolutionary implications. Proc Natl Acad Sci U S A 86(23):9412-6.

Wu, X. W., et al.

1992 Two independent mutational events in the loss of urate oxidase during hominoid evolution. J Mol Evol 34(1):78-84.

Yeldandi, A. V., et al.

1991 Molecular evolution of the urate oxidase-encoding gene in hominoid primates: nonsense mutations. Gene 109(2):281-4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- figures1.jpeg
- genemetadata.xlsx
- glud1glud2discriminatingsites.xlsx
- gluddiscsitepermutationnull.xlsx
- gludlocusbitscoredeltanoiseanalysisv3.xlsx
- gludpermutationnulldistribution10k.xlsx
- pa1spotcheckaggregatev2.xlsx
- paverificationlog.xlsx
- presenceabsencematrixverified.xlsx
- speciesmetadata.xlsx