Integrated and Evolving Genomic Landscapes:
Divergence, Selection and Functionalization Across Primate Lineages

Shrihith Golugula[1], Grant Merrigan[1], Rob J. Kulathinal, PhD[1]

Institute for Genomics and Evolutionary Medicine, Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA, USA

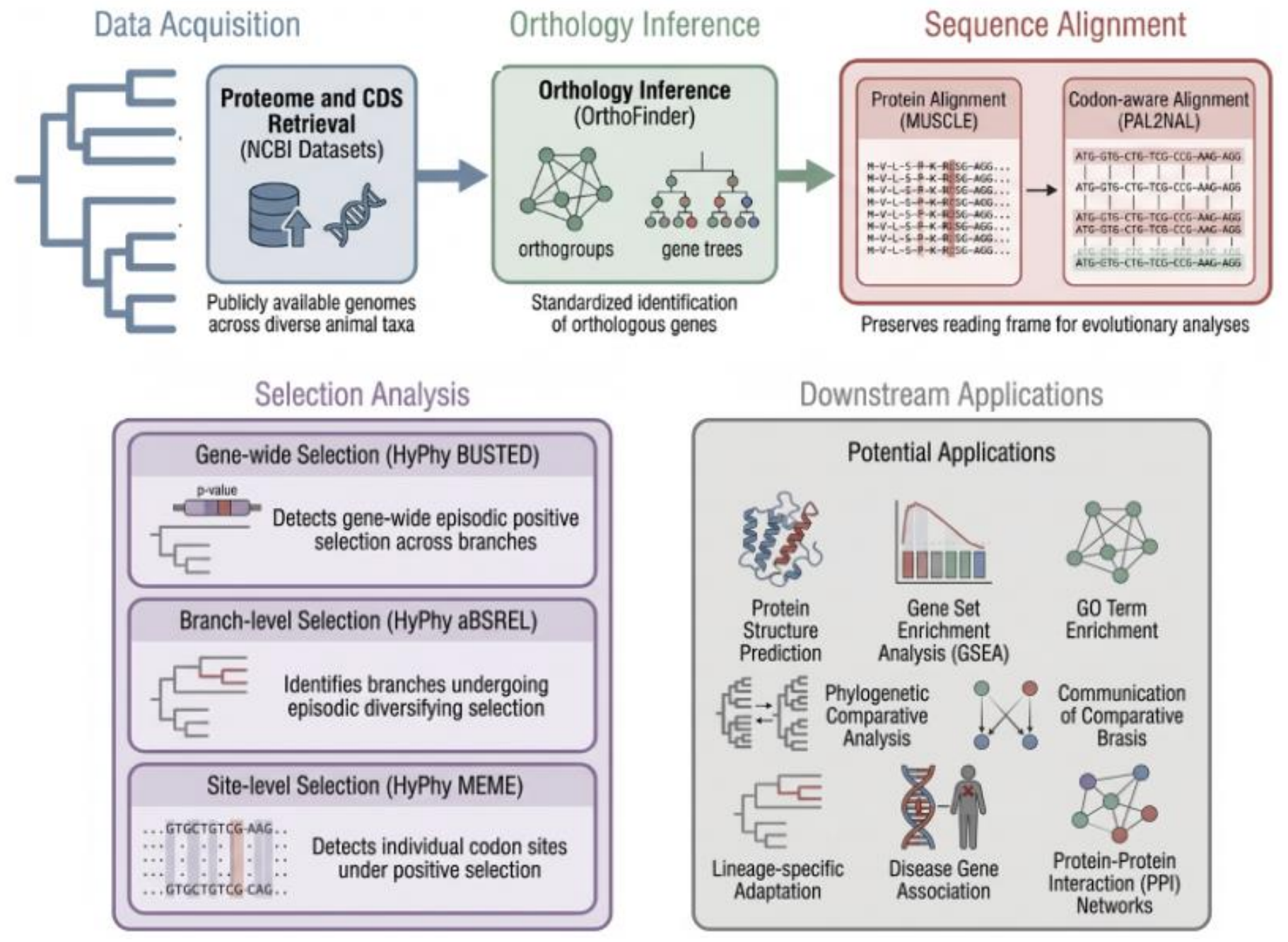
KULATHINAL LAB
EVOLUTIONARY GENOMICS

Abstract

Both positive Darwinian selection on conserved orthologs and the functionalization of new gene duplicates are major drivers of evolutionary innovation. Here, we utilize a comprehensive comparative genomic framework to investigate patterns of protein divergence and selection in orthologous genes as well as tissue diversification of duplicated gene families across eight representative primate genomes. We performed codon-aware alignments and phylogeny-informed selection analyses on 13,074 high-confidence 1:1 orthologs using HyPhy to detect gene-wide, branch-wide, and site-specific signals of protein evolution. Selection analyses revealed ample evidence of episodic diversifying selection across each of the primate lineages and functional enrichment tests highlight the overrepresentation of adaptively evolving genes in immune, reproductive, and sensory pathways. We then complemented this study of protein divergence and adaptation by examining functionalization patterns of gene duplicates in these lineages through the analysis of tissue-specific expression shifts in gene families exhibiting 1:2:1 copy number changes. Gene tree species tree reconciliation uncovered 1,104 duplication events across the primate phylogeny and, when mapped to a common set of tissues, a trend towards neofunctionalization was observed. Together, these results provide a divergence landscape in primates where positive selection is pervasive across the primate proteins and where gene duplication provides a substrate for innovation with selection driving functional diversification across lineages.

Analysis pipelines

Figure 1. Primate ortholog divergence and selection analysis pipeline



Data Acquisition
Proteome and CDS Retrieval (NCBI Datasets)
Publicly available genomes across diverse animal taxa

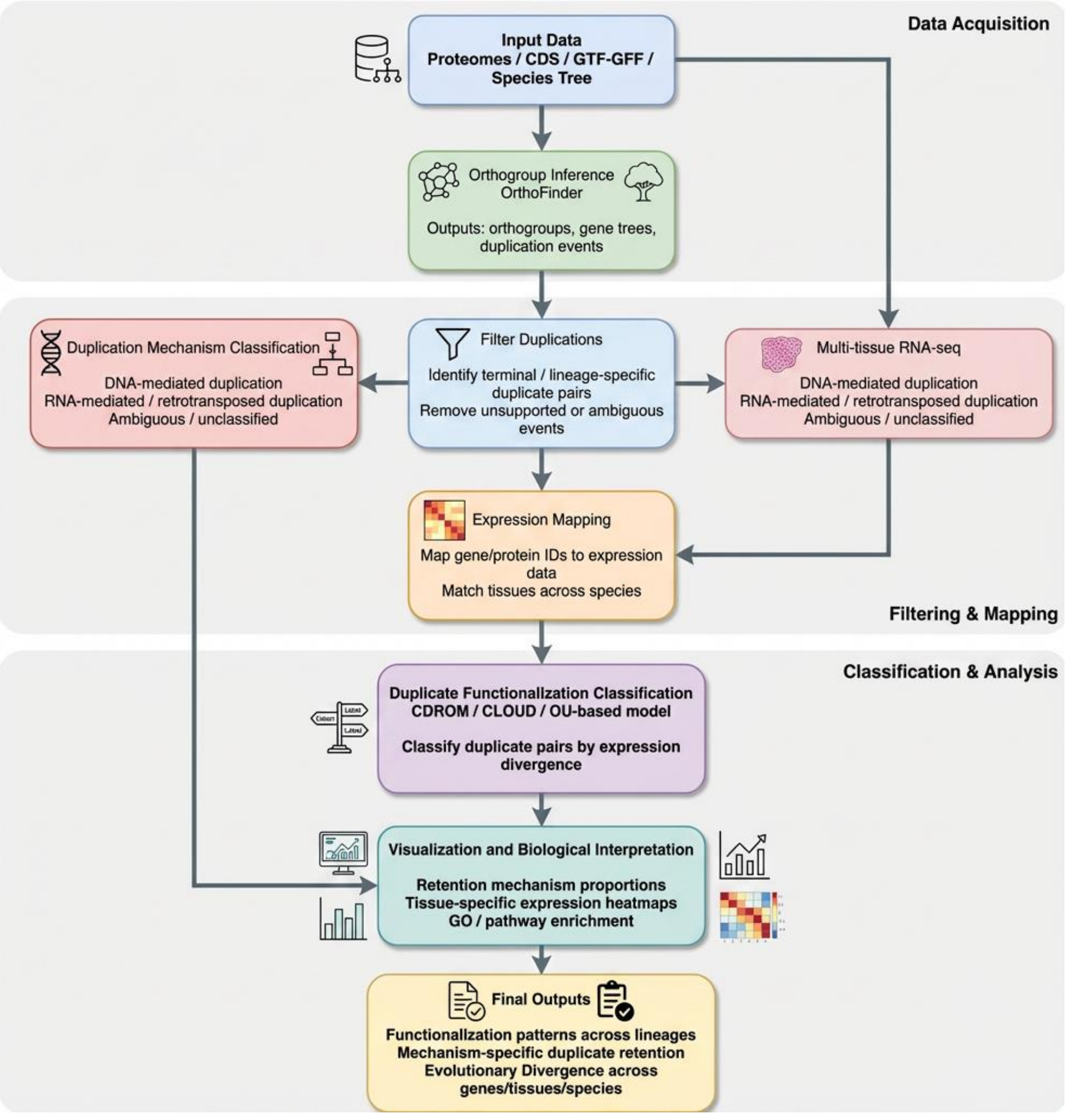
Orthology Inference
Orthology Inference (OrthoFinder)
orthogroups
gene trees
Standardized identification of orthologous genes

Sequence Alignment
Protein Alignment (MUSCLE)
Codon-aware Alignment (PAL2NAL)
Preserves reading frame for evolutionary analyses

Selection Analysis
Gene-wide Selection (HyPhy BUSTED)
Detects gene-wide episodic positive selection across branches
Branch-level Selection (HyPhy aBSREL)
Identifies branches undergoing episodic diversifying selection
Site-level Selection (HyPhy MEME)
Detects individual codon sites under positive selection

Downstream Applications
Potential Applications
Protein Structure Prediction
Gene Set Enrichment Analysis (GSEA)
GO Term Enrichment
Phylogenetic Comparative Analysis
Communication of Comparative Brasis
Lineage-specific Adaptation
Disease Gene Association
Protein-Protein Interaction (PPI) Networks

Figure 2. Duplicate gene functionalization analysis pipeline



Data Acquisition
Input Data
Proteomes / CDS / GTF-GFF / Species Tree
Orthogroup Inference
OrthoFinder
Outputs: orthogroups, gene trees, duplication events

Filtering & Mapping
Duplication Mechanism Classification
DNA-mediated duplication
RNA-mediated / retrotransposed duplication
Ambiguous / unclassified
Filter Duplications
Identify terminal / lineage-specific duplicate pairs
Remove unsupported or ambiguous events
Multi-tissue RNA-seq
DNA-mediated duplication
RNA-mediated / retrotransposed duplication
Ambiguous / unclassified
Expression Mapping
Map gene/protein IDs to expression data
Match tissues across species

Classification & Analysis
Duplicate Functionalization Classification
CDROM / CLOUD / OU-based model
Classify duplicate pairs by expression divergence
Visualization and Biological Interpretation
Retention mechanism proportions
Tissue-specific expression heatmaps
GO / pathway enrichment
Final Outputs
Functionalization patterns across lineages
Mechanism-specific duplicate retention
Evolutionary Divergence across genes/tissues/species

Episodic diversifying selection in 1:1 primate orthologs

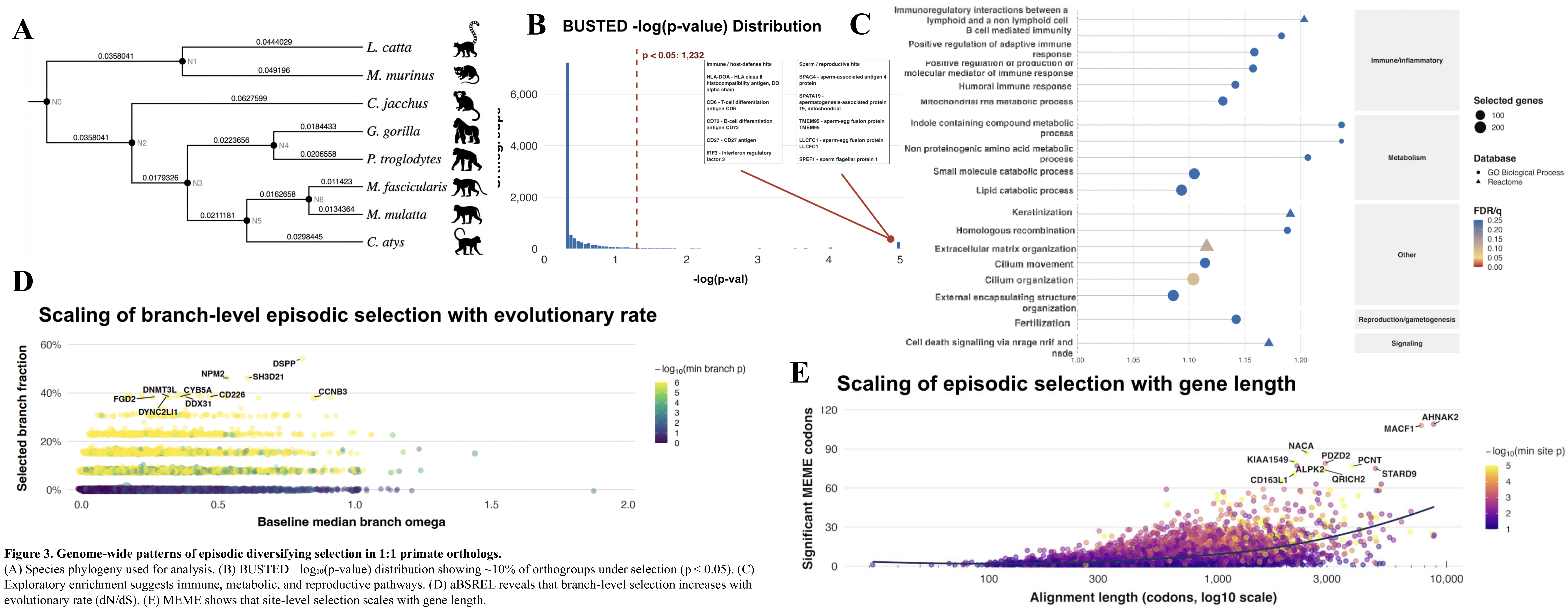


Figure 3. Genome-wide patterns of episodic diversifying selection in 1:1 primate orthologs.
(A) Species phylogeny used for analysis. (B) BUSTED −log₁₀(p-value) distribution showing ~10% of orthogroups under selection ($p < 0.05$). (C) Exploratory enrichment suggests immune, metabolic, and reproductive pathways. (D) aBSREL reveals that branch-level selection increases with evolutionary rate (dN/dS). (E) MEME shows that site-level selection scales with gene length.

Functionalization of primate duplicate genes

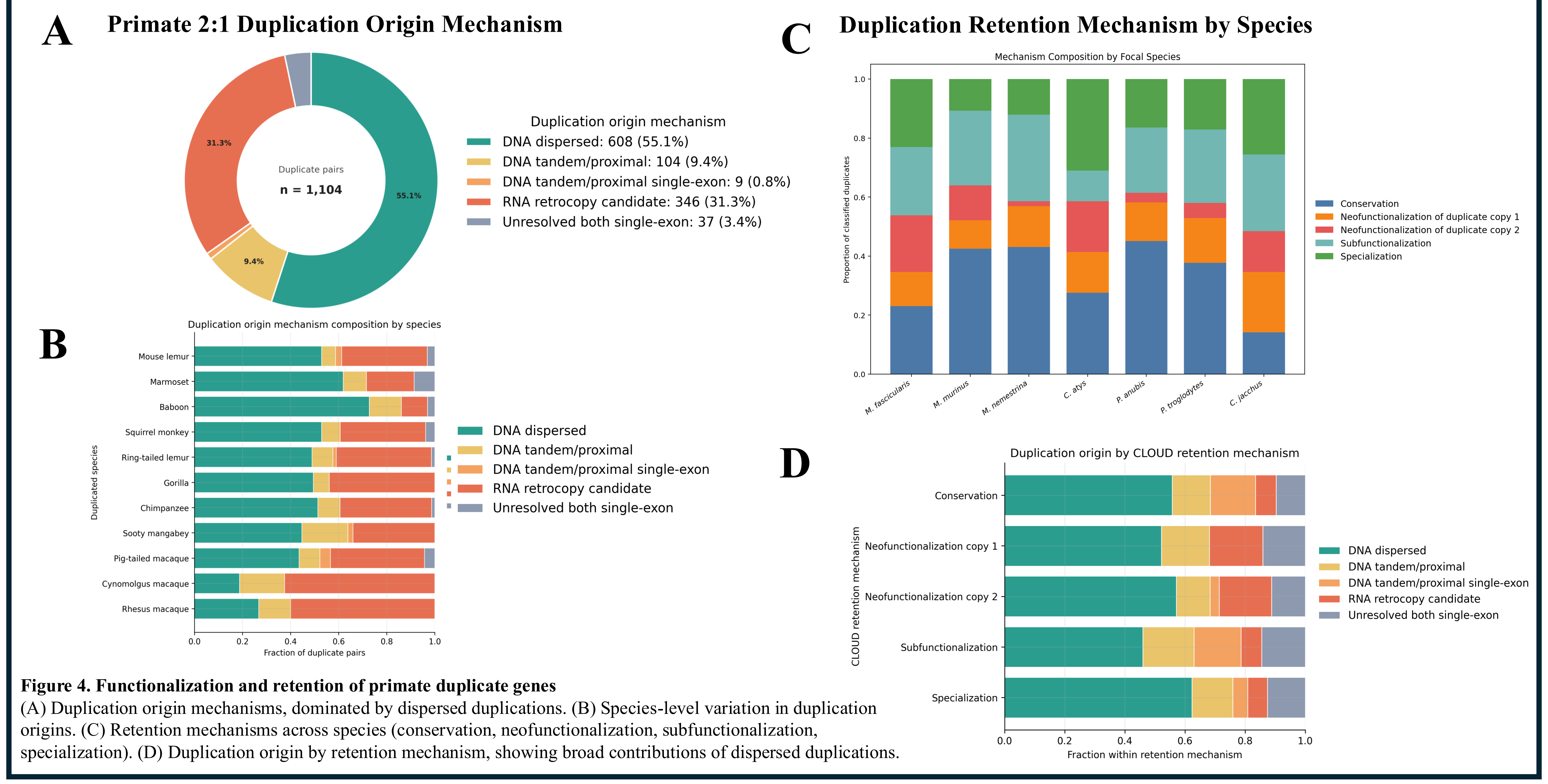


Figure 4. Functionalization and retention of primate duplicate genes
(A) Duplication origin mechanisms, dominated by dispersed duplications. (B) Species-level variation in duplication origins. (C) Retention mechanisms across species (conservation, neofunctionalization, subfunctionalization, specialization). (D) Duplication origin by retention mechanism, showing broad contributions of dispersed duplications.

Summary

1. Genome-wide selection landscape
Most primate orthologs are under purifying selection; ~10% show episodic diversification. Signals scale with dN/dS and gene length, with enrichment in immune, signaling, and reproductive pathways.

2. Duplication-driven innovation
~1,100 duplications (mainly dispersed) show diverse fates (conservation, neo-/subfunctionalization, specialization) indicating widespread divergence. Together, selection and duplication drive primate innovation.

Future directions

1. PhyloDIVaS: Integrate the divergence and selection analysis pipeline into a unified, scalable workflow for genome-wide evolutionary analysis. The web tool aims to increase access to comparative genomic analysis to experience-levels.

2. DupliCA: A modular web-based platform for gene duplication analysis that enables users to connect tools into reproducible pipelines for identifying duplicate genes, classifying duplication mechanisms, integrating expression data, estimating evolutionary divergence, and visualizing duplicate-gene retention patterns.

Acknowledgements

We would like to thank the **CodePhylo team** for their dedication and contributions to the development of the phyloDIVaS back-end pipeline used for the divergence and selection analysis (James McMichael, Meira Galapo, Hadyn Minnigh, Darien Sanith, George Vekios, Daniel Gillespie, Nick Mulvanerty, Tracy Lan, and Grant Merrigan).