

Computational Pipeline and Student Instructions

Course-Based Undergraduate Research Experience (CURE) | Animal Development
Susanta K. Behura | Proposed instructional resource

Purpose and scientific context

This guide introduces students to a published computational strategy for examining evolutionary divergence of duplicated (paralogous) genes. The original GitHub pipeline analyzes mouse gene families by combining coding sequences, Ensembl paralog annotations, gene expression, phylogenetic trees, and Bayesian model comparisons. In the proposed Animal Development CURE, students will first reproduce a guided mouse example and then extend the research question to expression conservation of orthologs and paralogs in mammalian embryos. The cross-species extension is a proposed instructional adaptation, not an existing feature of the linked script.

Primary source materials

Source code: [Behura Lab: Pipeline to Model Evolution of Paralog Expression](#) (17 numbered steps; accessed October 2026).

Research publication: Islam M, Behura SK. Role of paralogs in the sex-bias transcriptional and metabolic regulation of the brain-placental axis in mice. *Placenta*. 2024;145:143-150. DOI: 10.1016/j.placenta.2023.12.019.

Prerequisites and computing requirements

Introductory genetics, embryonic development and basic familiarity with files and spreadsheets. No prior Unix experience is required. Students need internet access and a Unix/Linux shell (local computer or institutional cluster). The GitHub source lists **wget**, **awk**, **sed**, **sort**, **Clustal Omega (clustalo)**, **Phyx (pxnj, pxrr, pxt2nex)**, **R with the ape package**, and **BayesTraitsV4**. Its module load commands assume an environment-module system and must be adapted to the local installation. Confirm access and licensing for BayesTraits before teaching.

Input data and research resources

Original mouse workflow: Ensembl mouse coding-sequence FASTA (GRCm39 in the script), Ensembl BioMart paralog annotations, and a prepared expression input named `expr_data`. The published example uses mouse day-15 placenta and fetal-brain RNA-seq and Ensembl homology annotations; it is a fetal tissue example, not a multi-species preimplantation embryo dataset.

Proposed CURE extension: obtain embryo expression datasets from EmExplorer, DevOmics, dbEmbryo, TEDD, EMBL-EBI Single Cell Expression Atlas, and/or NCBI GEO, depending on accessibility and comparability. Obtain ortholog and paralog relationships and stable gene IDs from Ensembl. Record species, developmental stage, assay, tissue/cell type, accession, normalization method and genome/annotation release. Match biologically comparable stages or cell types before cross-species comparisons.

Original GitHub workflow: 17 steps grouped for teaching

Steps	What students examine
1-5	Download mouse CDS; clean FASTA headers; convert FASTA to tabular form; select longest CDS per gene.
6-7	Retrieve/filter BioMart paralog relationships (script uses >25% identity in both relevant fields); link paralogs to CDS.
8-11	Integrate expression values; retain groups with at least five expressed genes; assign group/outgroup identifiers; generate sequence and expression files per paralog group.
12-14	Align sequences with Clustal Omega; construct and root trees with Phyx; convert to NEXUS and process trees with R ape.
15-17	Run BayesTraits model comparisons; collect marginal likelihoods; calculate log Bayes factors; flag groups with strong model-comparison evidence.

Suggested guided student activities

Activity A - Reproduce and inspect: Download the GitHub script; annotate its 17 steps, identify all inputs and outputs, and construct a workflow diagram. Run a small instructor-tested mouse example and verify the number of genes and paralog groups retained at each filtering stage.

Activity B - Gene-family investigation: Each team selects one embryonic-development gene family, obtains Ensembl homology annotations, retrieves suitable public expression data, checks stable gene identifiers and missing values, and visualizes expression by stage and species.

Activity C - Hypothesis testing: Compare within-family expression similarities for orthologous and paralogous gene pairs, controlling as feasible for species, developmental stage and cell type. For selected paralog groups, relate expression patterns to phylogeny and compare alternative evolutionary models. Distinguish expression conservation from direct proof of functional equivalence or adaptive selection.

Activity D - Reporting: Submit an analysis log, source-data citations, versioned scripts, a gene-family diagram, expression plots, a model-comparison summary where appropriate, and a short interpretation describing alternative explanations and limitations.

Important implementation notes and checks

The public script is a **research workflow outline rather than a fully turnkey student package**. Its BioMart download command contains an ellipsis placeholder ([http://www.ensembl.org/biomart/...](http://www.ensembl.org/biomart/)); instructors must supply a working export/query and confirm column ordering. The input expression-table format, outgroup mappings, and aggregation of BayesTraits marginal-likelihood outputs must be documented with example files before classroom execution. Check gene identifiers, delimiter assumptions, command versions, rooting choices, and whether output filenames match downstream commands.

The GitHub script labels the second BayesTraits run with `TestCorrel` and uses a log Bayes factor threshold of **>10** for strong evidence; the 2024 paper describes random-walk versus directional models and reports an evidence threshold of **>2** for its comparison. These are not interchangeable analyses or decision thresholds. The instructor should verify the intended BayesTraits commands and model definitions for the specific exercise rather than presenting them as equivalent. Neither correlation with phylogeny nor expression divergence alone establishes natural selection or neofunctionalization.

Expected student deliverables and quality assurance

Each group will produce (1) a dataset provenance/metadata sheet, (2) an Ensembl homology table with release/version, (3) reproducible code and software versions, (4) expression plots and phylogenetic analyses, and (5) a research report or presentation. Students will check for duplicate IDs, missing annotations, sample/stage comparability, expression normalization differences, and reproducibility of intermediate files. Instructor review at each milestone will precede interpretation.

Suggested public resources

Ensembl: <https://www.ensembl.org/> | GEO: <https://www.ncbi.nlm.nih.gov/geo/> | EMBL-EBI Single Cell Expression Atlas: <https://www.ebi.ac.uk/gxa/sc/home> | EmExplorer: <https://ngdc.cncb.ac.cn/databasecommons/database/id/6722> | DevOmics: <https://ngdc.cncb.ac.cn/databasecommons/database/id/7735> | dbEmbryo: <https://pubmed.ncbi.nlm.nih.gov/35977841/> | TEDD: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9825605/>