

# Transcriptome profiling of intra-snail stages of the liver fluke *Fasciola hepatica* reveals key mediators underlying parasite development and interaction with the host (Langleib et al., 2026)

## Overview

This repository accompanies the manuscript " *Transcriptome profiling of intra-snail stages of the liver fluke *Fasciola hepatica* reveals key mediators underlying parasite development and interaction with the host*" (Langleib et al., 2025). It centralises the trimmed reads, intermediate objects and code used to generate the figures, tables and supplementary analyses described in the paper. Only files up to 100 MB are tracked in Git to keep the repository lightweight.

Additional raw and processed data layers are archived on Zenodo: DOI:<https://doi.org/10.5281/zenodo.17204476>.

## Repository layout

- **data/** – Sequencing adapters and the subset of raw reads shared in the repository. Full FASTQ collections are referenced in the manuscript and Zenodo record.
- **notebooks/** – Jupyter notebooks that reproduce each table, figure and exploratory analysis. The first markdown cell of every notebook documents its intent, scope and execution environment.
- **scripts/** – Stand-alone Python and R scripts used throughout the pipeline (read trimming, mapping, expression parsing, enrichment analyses, etc.).
- **results/** – Output tables, plots and intermediate artefacts generated by the notebooks and scripts.

## Analysis roadmap

1. **Read preparation** – Adapter trimming, quality control and alignment against the *F. hepatica* Liverpool genome, with additional remapping of unmapped reads to *Lymnaea* hosts.
2. **Transcript assembly & quantification** – *StringTie* assemblies, *Trinity* transcriptome builds and expression quantification with *kallisto/sleuth* at transcript and gene level, consolidating canonical and putative novel CDS models.
3. **Expression filtering & summarisation** – CPM-based filtering to retain robust signal per biological condition, generation of multi-condition expression matrices and exploratory statistics (MDS, PCA, housekeeping candidates).
4. **Differential expression** – Pairwise and lifecycle-wide contrasts using *edgeR* with multiple QC checks, volcano plots and custom DEG clustering workflows.
5. **Module and comparative analyses** – Gene clustering, orthogroup classification, cross-species comparisons (including *F. gigantica* and *S. mansoni*), and assessment of stable versus variable genes.
6. **Functional interpretation** – GO/KEGG enrichment, KEGG pathway visualisation, BRITE category tallies, custom annotation parsing and remote homology searches to contextualise DEG clusters.

## Notebook catalogue

### Pre-processing, mapping and quantification

- `notebooks/001_mapping_unmapped_reads_to_lymnaea.ipynb` – Remaps *F. hepatica* reads that failed to align to the parasite genome against two *Lymnaea* reference genomes for contamination assessment.
- `notebooks/001_parsing_mapping_data_genelevel_full_cycle.ipynb` – R-based notebook that aggregates *kallisto* quantifications to gene level across the full lifecycle, applies condition-aware filters and prepares transcript-to-gene mappings including putative novel transcripts.
- `notebooks/006_running_all_pairwise_differential_expression_analyses_edgeR.ipynb` – Automates exhaustive pairwise *edgeR* analyses by reusing Degust logic while enforcing CPM thresholds per group.
- `notebooks/007_mapping_and_parsing_smansoni_bulk_RNAseq_from_Wang_and_Collins_studies.ipynb` – Downloads, maps and parses *Schistosoma mansoni* developmental bulk RNA-Seq datasets for comparative purposes.
- `notebooks/hisat2_to_get_logouts_in_lbc8.ipynb` – Regenerates HiSat2 alignment logs on LBC8 to obtain per-sample mapping summaries.

### Expression filtering, summarisation and QC

- `notebooks/000_filtering_expression_data_fhepatica.ipynb` – Applies CPM  $\geq 1$  per condition filtering to define the working *F. hepatica* gene set while retaining canonical and putative novel CDS.
- `notebooks/001_mds_fhepatica_samples.ipynb` – Performs MDS analyses on the filtered gene set to visualise lifecycle relationships.
- `notebooks/analisis_descriptivos_sobre_estadios_novedosos.ipynb` – Collects descriptive statistics for novel intra-snail stages, including expression summaries and manual curation checkpoints.
- `notebooks/checkeo_respecto_a_DEGs_y_false_positive_rate.ipynb` – Evaluates DEG false-positive behaviour arising from sparse CPM profiles and adjusts thresholds accordingly.
- `notebooks/defining_housekeeping_genes.ipynb` – Benchmarks multiple criteria to nominate housekeeping genes across *F. hepatica*, *F. gigantica* and *S. mansoni*.
- `notebooks/plotting_stable_and_variable_genes.ipynb` – Visualises trends uncovered in `defining_housekeeping_genes.ipynb`, focusing on stability metrics and expression distributions.
- `notebooks/recalculating_DEGs.ipynb` – Revisits the DEG detection pipeline to correct earlier Degust-derived inconsistencies.

### Differential expression visualisation and summarisation

- `notebooks/004_DEG_results_pairwise_lifecycle.ipynb` – Explores lifecycle DEG tables through KEGG KO annotations, focusing on sequential stage transitions.
- `notebooks/004_volcano_plots_transitions_MIR-IS_IS15-IS30.ipynb` – Generates publication-ready volcano plots for MIR→IS and IS15→IS30 comparisons.
- `notebooks/005_plotting_DEGs_clusters_logFC_1.5.ipynb` and `notebooks/Ploteando_clusters_DEGs_logFC_1.5.ipynb` – Produce polished

heatmaps of DEG clusters at  $|\log FC| \geq 1.5$ .

- `notebooks/DEG_analyses_on_lengths_and_BLAST_hits.ipynb` – Quantifies DEG length distributions and BLAST matches, controlling for sequence length biases.

## Clustering and module exploration

- `notebooks/004_running_clust.ipynb` – Reruns `clust` module detection and reproduces cluster plots.
- `notebooks/alternatives_for_clustering_analysis.ipynb` – Tests variance-based stratification prior to clustering to contrast module structures.
- `notebooks/clustering_DEGs.ipynb` – Clusters DEGs combining canonical genes and putative novel CDS, with emphasis on module interpretation.
- `notebooks/clustering_of_genes.ipynb` and `notebooks/clustering_of_genes.different_sample_set.ipynb` – Compare multiple clustering algorithms (hierarchical, k-means, SOMs) on *F. hepatica* and *F. gigantica* datasets.
- `notebooks/cluster_comparison_based_on_KOs.ipynb` – Aligns DEG clusters with KEGG annotations to assess concordance across analyses.
- `notebooks/functional_analysis_of_clusters.ipynb` – Performs GO/KEGG enrichment on clusters, combining canonical and novel annotations.
- `notebooks/orthogroup_analysis_clustered_degs_fasciolidae.ipynb` – Interrogates orthogroup composition within DEG clusters across fasciolid species.
- `notebooks/chiveando_sobre_posibles_valores_de_corte_borrar.ipynb` – Scratchpad notebook used to probe OG filtering thresholds (kept for reference).

## Comparative and evolutionary analyses

- `notebooks/005_searching_for_homology_between_fasciolidae_top_expressed_genes.ipynb` – Studies whether highly expressed genes share orthogroups between *F. hepatica* and *F. gigantica*.
- `notebooks/analyzing_smansonii_developmental_markers.ipynb` – Maps *S. mansoni* lifecycle markers onto fasciolid datasets for cross-species comparisons.
- `notebooks/retrieve_taxonomy_scope_of_homology_for_fasciolidae_genes.ipynb` – Retrieves taxonomic breadth of homology for fasciolid genes via Entrez.
- `notebooks/phenetic_distances_between_DEGs_and_stable_genes.ipynb` – Compares evolutionary distances between DEG sets and stable genes.
- `notebooks/008_obtaining_and_processing_cestoda_and_turbellaria_temporal.ipynb` – Downloads and processes Turbellaria/Cestoda proteomes for distant outgroup analyses.
- `notebooks/corriendo_búsqueda_homologos_remotos.ipynb` – Documents remote homology (HMM–HMM) searches requiring long runtimes.
- `notebooks/classify_orthogroups_of_fasciolidae.ipynb` – Classifies orthogroups containing fasciolid members and annotates their composition.

## Functional annotation and enrichment

- `notebooks/008_exploring_genes_high_logFC_avgExpr.ipynb` – Identifies genes with concurrent high logFC and average expression.
- `notebooks/contando_KEGG_brites.ipynb` – Tallies KEGG BRITE hierarchies for stable genes, DEG clusters and high-logFC subsets.
- `notebooks/categorias_funcionales_MIR_IS.ipynb` – Details KEGG/COG categories in MIR and intra-snail expression profiles.
- `notebooks/010_taxonomic_conservation_of_stable_genes_and_DEGs.ipynb` – Assesses conservation patterns of stable genes and DEGs across taxa.

## Script catalogue

---

### Python utilities

- `scripts/python/000_mapping_kallisto.py` – Automates *kallisto* indexing and quantification for paired-end libraries, creating the mapping directory structure as needed.
- `scripts/python/000_retrievieng_fgigantica_metadata.py` – Uses NCBI Entrez to retrieve *F. gigantica* sequencing metadata and exports run accession tables.
- `scripts/python/000_retrievieng_smansonii_metadata.py` – Mirrors the metadata retrieval pipeline for *S. mansoni* datasets.
- `scripts/python/000_trinity_run.py` – Wraps *Trinity* genome-guided transcriptome assemblies for all samples on the LBC8 server.
- `scripts/python/001_mapping_hisat2.py` – Builds HiSat2 indices and maps libraries against the *F. hepatica* genome in batch mode.
- `scripts/python/002_sam2bam.py` – Converts SAM alignments to BAM format using `samtools`.
- `scripts/python/003_stringtie_assembly.py` – Runs *StringTie* assembly/merge steps and expression estimation through helper routines in `my_functions.py`.
- `scripts/python/BLASTPing_all_clusters_fhepatica_vs_fgigantica_t4.py` – Executes BLASTP comparisons between cluster members of *F. hepatica* and *F. gigantica*.
- `scripts/python/align_demonstrative_families.py` – Aligns selected orthogroups with MAFFT (L-INS-i mode) for illustrative alignments.
- `scripts/python/mapping_missing_smansonii_reads.py` – Maps previously unmapped *S. mansoni* reads back to the genome to fill coverage gaps.
- `scripts/python/my_functions.py` – Shared helpers for directory creation, *StringTie* operations and *kallisto* quantification.
- `scripts/python/parsing_KEGG_enrichment_tables.py` – Parses KEGG enrichment outputs, fetching pathway and module descriptions via the KEGG REST API.
- `scripts/python/parsing_functional_annot.py` – Merges EggNOG annotations, retrieves KO metadata in parallel and exports consolidated annotation tables.

### R utilities

#### Data parsing and filtering

- `scripts/R/001_trimming_and_fastqc.R` – Coordinates adapter trimming and FASTQC reporting across libraries.
- `scripts/R/reading_mapping_data.R` – Loads mapping outputs into tidy structures for downstream parsing.
- `scripts/R/parsing_mapping_data.R` – Aggregates transcript-level *kallisto* results, labelling canonical and putative novel transcripts.
- `scripts/R/parsing_mapping_data_genelevel.R` – Extends the parser to gene-level summaries via transcript-to-gene dictionaries.
- `scripts/R/parsing_mapping_data_genelevel_full_cycle.R` – Full lifecycle parser consolidating all stages, including novel transcript class filtering.
- `scripts/R/parsing_mapping_data_genelevel_fgigantica.R` – *F. gigantica*-specific expression parsing.
- `scripts/R/parsing_mapping_data_genelevel_is_miracidia.R` – Focused parser for miracidia and intra-snail subsets.
- `scripts/R/parsing_mapping_data_genelevel_smansonii.R` – Equivalent parser for *S. mansoni* datasets.

- `scripts/R/filtering_orfs_transcript_level_intrasnail.R` and `scripts/R/filtering_orfs.R` – Filter predicted ORFs from putative novel transcripts at transcript level.
- `scripts/R/000_parsing_mapping_data_genelevel_full_cycle.R` – Notebook-equivalent script consolidating gene-level counts across stages.
- `scripts/R/obtaining_data_module_clustering.R` – Prepares expression matrices for module clustering, including novel CDS filtering.

#### Differential expression and summarisation

- `scripts/R/differential_expression.R` – Runs *edgeR*-based DEG analyses for lifecycle comparisons.
- `scripts/R/differential_expression_tables.R` and `scripts/R/differential_expression_tables_HE.R` – Generate tidy DEG summary tables for intra-snail stages and highly expressed subsets.
- `scripts/R/subsetting_lists_for_differential_expression.R` – Builds curated gene lists for DEG contrasts.
- `scripts/R/comparing_genes_in_DE_intrasnail.R` – Checks overlap between intra-snail DEG contrasts to identify shared signals.
- `scripts/R/comparing_HE_miracidia_intrasnails.R` – Compares highly expressed gene sets between miracidia and intra-snail stages.
- `scripts/R/highly_expressed_transcript_analyses.R` – Explores expression profiles of highly expressed transcripts across stages.

#### Clustering and module interrogation

- `scripts/R/clust_interrogating_clusters_intra-fhepatica.R` – Explores `clust` output for the full lifecycle dataset.
- `scripts/R/clust_module_plotting.R` – Replots `clust` modules with corrected axis labels and optional restriction to novel CDS.
- `scripts/R/clust_subclustering_data_retrieval.R` – Retrieves data to perform subclustering on selected modules.
- `scripts/R/plot_clust_varying_parameter_results.R` – Visualises how `clust` parameter choices affect module counts.
- `scripts/R/plot_clust_multigene_family_dimensions.R` – Summarises multigene family composition and novel CDS content per cluster.

#### Functional annotation and enrichment

- `scripts/R/000_GO_enrichment_analysis.R` and `scripts/R/000_GO_enrichment_analysis_HE.R` – Perform GO enrichment with `topGO` / `ViSEAGO` for general and highly expressed gene sets.
- `scripts/R/KEGG_enrichment_analysis_DE_gProfileR.R` and `scripts/R/KEGG_enrichment_analysis_HE_gProfileR.R` – Run KEGG enrichment via `gProfiler2`.
- `scripts/R/pathview_plotting.R` – Projects log2 fold-changes onto KEGG pathways using `pathview`.
- `scripts/R/plotting_mds.R` and `scripts/R/plotting_mds_genelevel.R` – Produce MDS plots from transcript- and gene-level matrices.

#### Orthology and comparative genomics

- `scripts/R/000_renaming_trematode_proteomes.R` – Renames trematode proteomes with harmonised tags for orthology inference.
- `scripts/R/reading_orthogroups.R` – Imports orthogroup assignments for downstream interrogation.
- `scripts/R/classify_orthogroups_of_fasciolidae.R` – Classifies fasciolid orthogroups (1:1, 1:many, many:many) and annotates descriptions.
- `scripts/R/BLASTping_all_clusters_fhepatica_vs_fgigantica_t4.R` – BLASTP comparisons across fasciolid cluster members with novel CDS handling.
- `scripts/R/searching_for_homology_between_fasciolidae_top_expressed_genes.R` – Tests orthogroup overlap for top expressed genes in both fasciolids.
- `scripts/R/searching_homology_HE_miracidia_fascilas.R` – Probes whether *F. gigantica* highly expressed genes have expressed homologues in *F. hepatica*.
- `scripts/R/000_novel_transcript_analyses.R` – Characterises StringTie-derived novel transcripts, predicted ORFs and length distributions.

## Contact

For questions regarding this study or the associated data, please contact MSc. Mauricio Langleib ([mauricio.langleib@gmail.com](mailto:mauricio.langleib@gmail.com), [mlangleib@higiene.edu.uy](mailto:mlangleib@higiene.edu.uy), [mlangleib@pasteur.edu.uy](mailto:mlangleib@pasteur.edu.uy)), Dr. Andrés Iriarte ([airiartheo@higiene.edu.uy](mailto:airiartheo@higiene.edu.uy)) or Dr. Jose F. Tort ([jtort@fmed.com.uy](mailto:jtort@fmed.com.uy)).

[https://github.com/mauriciolangleib/fhepatica\\_snailseq/](https://github.com/mauriciolangleib/fhepatica_snailseq/)