EE282 Final Project Report *

Valentina Peña *Nutritional Physiology Lab, UC Irvine*

METHODS

Experimental design

Liver paired-end raw sequencing files for all biological replicates of *Xiphister mucosus* and *Phytichthys chirus* were obtained from the original study [1] using SRA-tools v3.0.0 and the prefetch tool from the SRA database. The downloaded .sra files were converted to .fastq format using fasterq-dump, organizing the paired-end reads. Following extraction, the .fastq files were compressed to .fastq.gz format. The Bash script used for downloading and processing these files is located at */code/ncbi_download_sequences_submit.sh*, and the raw reads are stored in */data/raw/fastq*.

Pre-assembly quality control

Quality reports for the raw read data were generated for each sample using FastQC [2] and aggregated into a single report with MultiQC [3]. These reports are located at: *data/raw/fastqc_results*. After inspecting the reports, adapter contamination and low-quality base calls were removed using Fastp [4], which produced individual quality reports for each sample, post-trimming. The trimming Bash script is located at *data/raw/fastq/fastp_job.sh*. The post-trimming reports and high-quality cleaned reads are stored in *output/reports/fastp* and *data/processed/trimmed_out*, respectively.

De novo transcriptome assembly and annotation

Following the approach of Herrera et al. (2022), transcriptome assemblies were conducted using Trinity v2.15.2 [5]. RNA-seq data for X.mucosus and P.chirus were processed separately to produce species-specific assemblies. Each assembly incorporated pooled RNA-seq data across experimental conditions to maximize transcript diversity. Trinity performed *in silico normalization* of reads to reduce systematic coverage and improve assembly performance. Batch jobs were submitted using custom scripts located at *code/scripts*. The script TrinityStats.pl was utilized to summarize basic metrics for the P.chirus and X.mucosus *de novo* assemblies which can be found in *output/reports*.

Prefixes 'pc' and 'xm' were used throughout the final project directory to distinguish subdirectories and files for P.chirus and X.mucosus, respectively. For example, Trinity outputs for P. chirus and X.mucosus are found in *data/processed/pc_trinity_out.Trinity.fasta.gz data/processed/xm_trinity_out.Trinity.fasta.gz*.

Transcriptome annotation was performed using Trinotate v4.0.2 [6]. Open reading frames (ORFs) were identified using TransDecoder v2.0.1 [7], and translated ORFs were aligned against the Swiss-Prot database using BLASTX. The best hits were annotated with Gene Ontology (GO) terms. Protein domains were identified with HMMER v3.1 and hmmscan against the Pfam-A

*Replication files are available on the author's Github account (http://github.com/ValentinaP-NYC). **Current version**: December 13, 2024; **Corresponding author**:valenp1@uci.edu.

database. Individual Trinotate reports were generated for each species and are located at *output/reports/pc_debugged_trinotate.xls.zip* and *output/reports/xm_debugged_trinotate.xls.zip*. The *.xls* files were too large and thus were not copied into the real_final_project_branch.

Transcript differential expression and enrichment analysis

Reads were aligned to the de novo assembled contigs using Bowtie2 [8], and transcript abundance was quantified with RSEM, enabling isoform- and gene-level quantification by clustering transcripts representing splicing isoforms [9]. Relative expression was reported as fragments per kilobase of transcript per million mapped reads (FPKM). Downstream analyses focused on the *pc_all* and *xm_all* matrix outputs, located in *data/processed/abundances*.

Differential expression analysis was conducted using the combined replicates *all.isoform.counts.matrix* file for each species to identify differentially expressed transcript isoforms across dietary conditions. Analysis was performed with EdgeR implemented with Trinity, using an FDR cutoff of 0.01, and transcripts with a fold change greater than 2.5 were classified as significantly differentially expressed [10]. Gene Ontology (GO) enrichment analysis of the differentially expressed transcript isoforms was performed using GOseq, accounting for gene length bias [11]. Biological sample quality was assessed with the Trinity PtR script, which generated PCA plots to evaluate clustering of individuals within each species across dietary conditions. The outputs are located in: *data/processed/edgeR/pc*, *data/processed/edgeR/xm*, and *data/processed/edgeR/PtR*. The code to conduct the analysis can be located at: *data/processed/edgeR/edgeR_script.R*.

RESULTS

Principle component analysis

For P.chirus, PC1 explains 72.57% of the total variance, followed by PC2 with 18.50% and PC3 with 8.93% (Fig.2A-B). A heatmap of the top 50 features contributing to variance in PC2 demonstrate clearer clustering of replicates within dietary conditions, based on isoform-level expression (Supplemtary material S1). This indicates that isoform expression in P. chirus was influenced by dietary conditions to a notable extent. Additional support for this can be noted in the total number of unique DE transcripts which is 1812

In X.mucosus, 37.61% of the total variance is explained by the first principal component (PC1), with PC2 accounting for 27.99%, and PC3 capturing an additional 17.88% (Fig.2C-D). Heatmaps of the top 50 features contributing to variance in PC1, PC2, and PC3, respectively, reveal that replicates within dietary conditions do not cluster distinctly. This observation suggests that isoform-level expression in X.mucosus is not strongly influenced by dietary condition. Across all three dietary conditions there were only 290 DE transcripts at the criteria (FDR < 0.01) and FC > 2.5.

Transcriptome assembly and annotation

Transcripts were subjected to annotation analysis by comparing with Nr, Nt, Pfam, Swiss-Prot, KEGG and GO databases. Trinotate result summary shows approximately 38% of unique genes in P.chirus and 43% of unique genes in X.mucosus were assigned Pfam IDs during annotation. Annotation revealed 6,172 Pfam domains in P. chirus and 5,369 in X. mucosus for unique genes Table 1). For total genes, P.chirus had 65,691 Pfam IDs, while X. mucosus had 55,925.

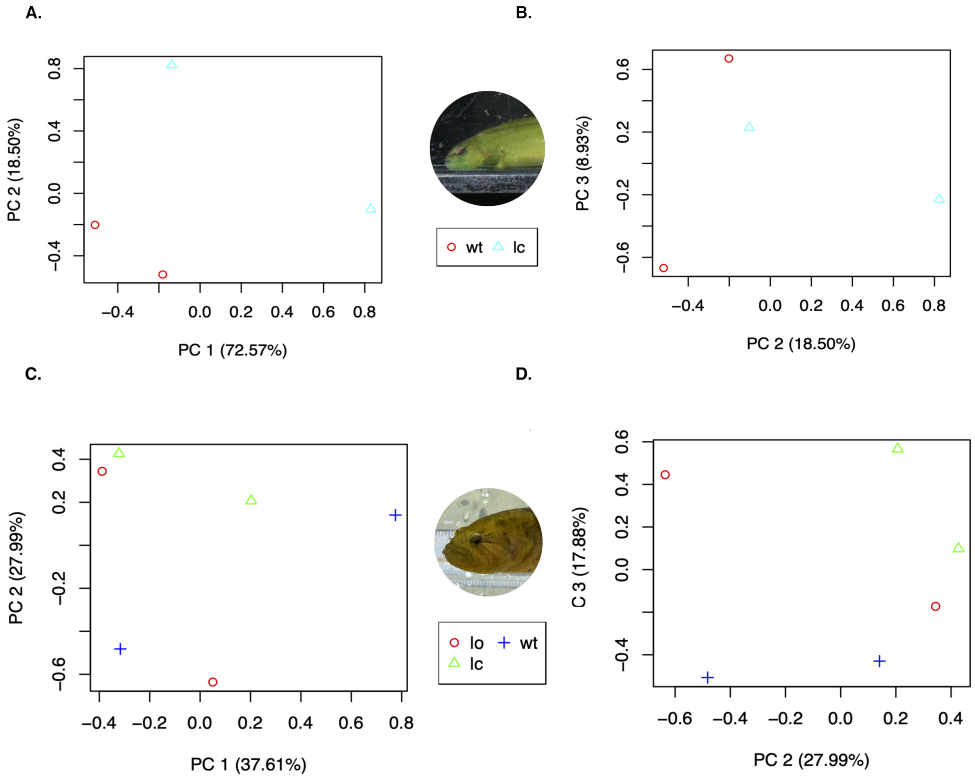

Figure 1: Principle components (PCA) plot for (A-B) P.chirus (n=4) and (C-D) X.mucosus (n=6). Replicates within a species are colored according to dietary treatment (lo = lab-omnivore; lc = lab-carnivore; wt = wild-type/no dietary intervention.) Lowly expressed transcripts (< 0.5 CPM) were not considered.

Species	P.chirus (Unique)	P.chirus (Total)	X.mucosus (Unique)	X.mucosus (Total)
Pfam	6172	65691	5369	55925
Genes	16126	30076	13007	23304
Transcripts	26181	49723	23402	42952
Proteins	28588	51160	25891	44524

Table 1: Pfam summarized Trinotate reporting. Annotated features (unique and total) for X.mucosus and P.chirus assemblies.

Transcript-level quantification and differential expression analysis

In this study, liver RNA-seq data from X. mucosus and P. chirus were analyzed to assess differential isoform-level expression in response to dietary conditions. The objective was to explore how diet-induced changes influence gene expression, focusing on alternative isoform usage. Transcript abundance was quantified using RSEM, and differential expression (DE) analysis was conducted in EdgeR with an FDR <0.01 and a fold change (FC) >2.5 across all pairwise comparisons.

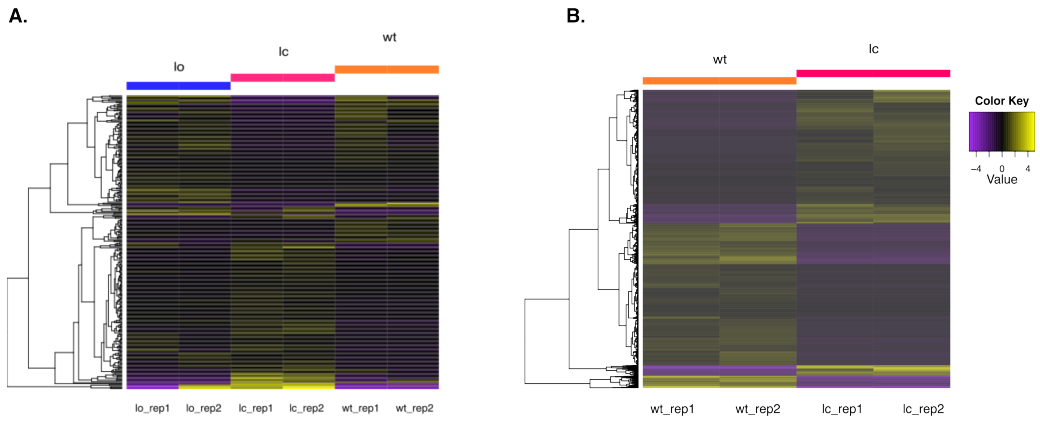

Figure 2: Liver transcript-level expression profiles for replicates of (A.) X.mucosus and (B.) P.chirus. Each row is a transcript isoforms and are clustered based on expression similarity in a dendrogram. Columns represent biological replicates of dietary conditions within a species (lo = lab-omnivore; lc = lab-carnivore; wt = wild-type/no dietary intervention). Lowly expressed transcripts (< 0.5 CPM) were not considered.

The differential expression analysis revealed distinct responses to diet in the two species. In P. chirus, a total of 1812 unique transcript isoforms were differentially expressed between wild-type and lab-carnivore individuals, indicating a substantial response to dietary changes. In contrast, X. mucosus exhibited fewer DE transcripts, with a total of 290 DE transcript isoforms across dietary conditions. Specifically, 127 transcript isoforms were differentially expressed when

comparing lab-carnivore to lab-omnivore-fed fish, and 182 transcript isoforms were DE between lab-carnivore and wild-type fish. Only 21 transcript isoforms showed significant differential expression between lab-omnivore and wild-type X.mucosus (Table 2).

Species	Comparison	Differentially Expressed Transcripts
P.chirus	Lab-Carnivore vs. Wild-Type	1812
X.mucosus	Lab-Carnivore vs. Lab-Omnivore	127
X.mucosus	Lab-Carnivore vs. Wild-Type	182
X.mucosus	Lab-Omnivore vs. Wild-Type	21

Table 2: EdgeR summarized DE transcripts output for comparisons across dietary condition within species at thresholds FDR <0.01 and FC > 2.5.

Enrichment analysis

To gain insights into the biological significance of the differentially expressed isoforms, Gene Ontology (GO) term enrichment analysis was performed using GOseq. This analysis annotated DE isoforms to biological, cellular, and molecular terms to uncover functional categories associated with the transcriptional response to diet in P.chirus and X.mucosus. GO term analysis of the annotated DE transcripts in wild-type P.chirus most were enriched in the "biological processes" category which included 49 transcripts in adaptive immune response (GO:0002250) term, 37 transcripts in immunoglobulin complex (GO:0019814) term (Fig 3A). Among the upregulated DE transcripts enriched in LC P.chirus, 67 were mapped to lipid metabolic processes (GO:0006629) term, 36 in transmembrane transporter activity (GO:0022857) term, and 29 in response to nutrient levels (GO:0031667)(Fig 3B).

Generally, fewer transcripts seemed to be differentially expressed across dietary condition for X. mucosus (Table 2). Between wild-type X.mucosus relative to LC individuals enriched, broadly, in immune cell activation and function (GO:0042110, GO:0046649, GO:0050776) (Fig4.A). For DE transcripts unregulated in LC X.mucosus, 49 mapped to ion binding (GO:0043167) molecular function, and 47 transcripts in catalytic activity (GO:0003824) biological processes (Fig.4B). Additionally, we performed GO enrichment analysis on a subset of the DE transcripts enriched in LC X.mucosus relative to LO X.mucosus (Fig.4C). For example, 24 transcripts upregulated in LC X.mucosus were involved in nitrogen compound metabolic processes (GO:0006807), 17 in organic cyclic compound metabolism (GO:1901360).

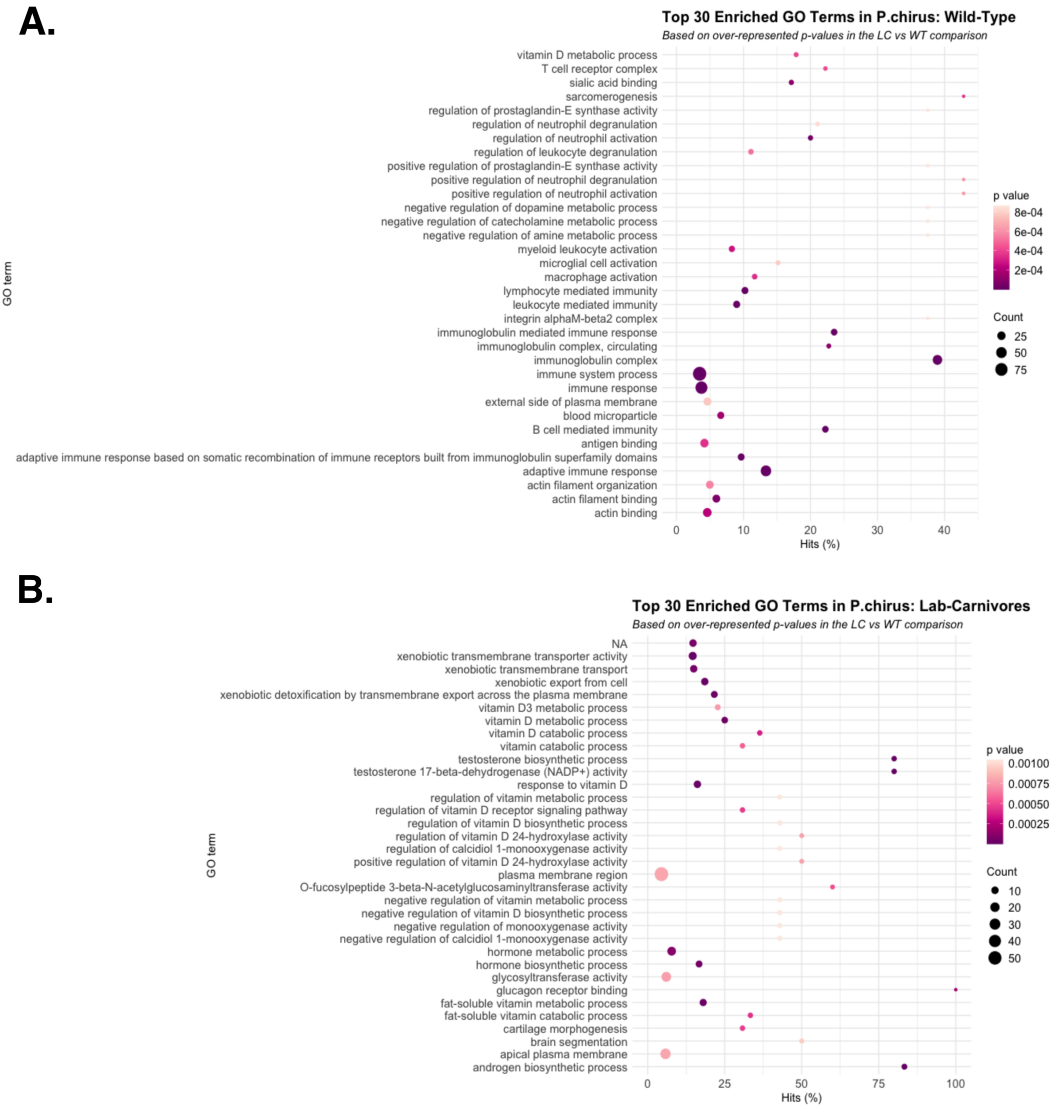

Figure 3: Gene Ontology (GO) term enrichment analysis. An overrepresented p value of < 0.01 was used to pick significantly enriched GO terms in P.chirus.

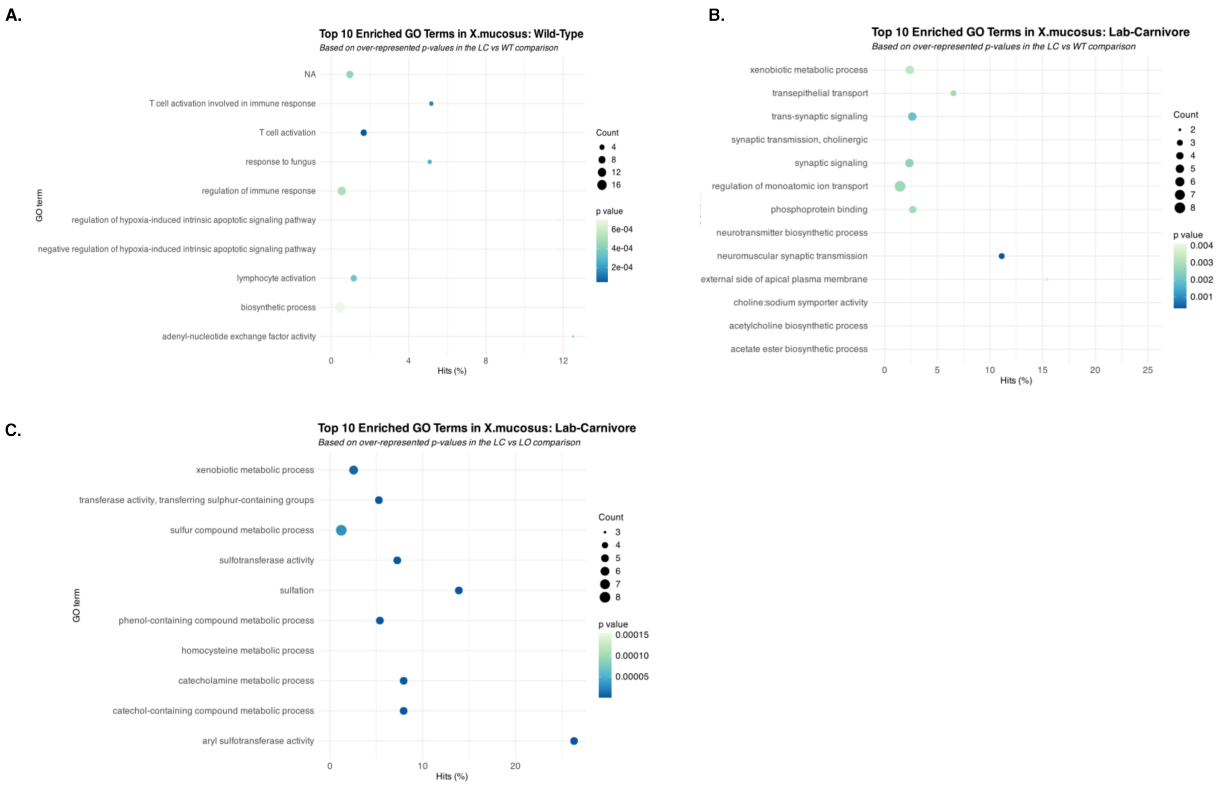

Figure 4: Gene Ontology (GO) term enrichment analysis. An overrepresented p value of < 0.01 was used to pick significantly enriched GO terms in X.mucosus.

DISCUSSION

RNA-Seq analyses have become a reliable method for assessing transcriptional responses to varying experimental conditions. Previous gene-level differential expression analysis in P.chirus revealed unique metabolic flexibility in response to dietary shifts, when compared to other intertidal Stichaeid species with different diets. Expanding on this framework, I explored isoform-level changes in both P.chirus and X.mucosus to determine whether these closely related species exhibited similar or distinct patterns, driven by their natural dietary habits.P.chirus, an omnivore, naturally consumes a diet rich in protein and carbohydrates, while X.mucosus, an herbivore, relies on an algae-based diet that is low in protein and fat but high in cellulose.

From the original study, when consuming the laboratory carnivore diet (LC), P.chirus exhibited enriched 13C and 15N signatures in the liver [1]. These stable isotopic data indicated that P.chirus was capable of assimilating more carbohydrates and proteins enriched in the LC diet compared to wild-type individuals that were not made to stray from their natural diets. Further comparison of X.mucosus under the wild-type and LC X.mucosus had upregulated transcripts related to protein metabolic processes and changes in transmembrane trafficking, suggesting metabolic switches to processing and transporting nutrients derived from increased protein intake.

The findings presented here also show that P.chirus exhibits a broader transcriptional response at the isoform-level in response to a LC diet, aligning with the earlier gene-level results. In both species, the wild-type individuals (P.chirus and X.mucosus) exhibited upregulated transcripts enriched for immune function and cell activation, likely reflecting the complexity of their natural environments, where exposure to parasites and microorganisms stimulates adaptive and innate immunity compared to laboratory conditions.

Limitations and future directions

A limitation of this study was the inability to accurately identify alternative splicing events in liver transcripts due to the constraints of short-read sequencing technology. Illumina short reads are limited in their ability to resolve complex splicing events, and as a result, the analysis presented here was limited to DE detection and GO enrichment. However, with the development of new genomic resources for P.chirus, it will be possible to map differentially expressed transcript isoforms and assess read coverage and exon structure for protein-coding genes. The integration of additional bioinformatics tools, such as long-read sequencing and splicing-aware software, will improve upon one of the original project goals of understanding alternative splicing and how it might influence dietary diversification in this species which exhibits high dietary flexibility under laboratory conditions.

References

Herrera, M.J., Heras, J. & German, D.P. Comparative transcriptomics reveal tissue-level specialization towards diet in prickleback fishes. J. Comp. Physiol. B 192, 275–295 (2022). https://doi.org/10.1007/s00360-021-01426-1.

Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). https://doi.org/10.1038/nmeth.1923.

Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890 (2018). https://doi.org/10.1093/bioinformatics/bty560.

Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). https://doi.org/10.1186/1471-2105-12-323.

Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011). https://doi.org/10.1038/nbt.1883.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010). https://doi.org/10.1093/bioinformatics/btp616.

Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016). https://doi.org/10.1093/bioinformatics/btw354.

Brian, H. & Papanicolaou, A. Transdecoder (Find Coding Regions Within Transcripts). Retrieved from http://transdecoder.github.io (n.d.).

Bryant, D.M. et al. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. Cell Rep. 18, 762–776 (2017). https://doi.org/10.1016/j.celrep.2017.01.002.

Young, M.D., Wakefield, M.J., Smyth, G.K. et al. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11, R14 (2010). https://doi.org/10.1186/gb-2010-11-2-r14.

Supplemental material

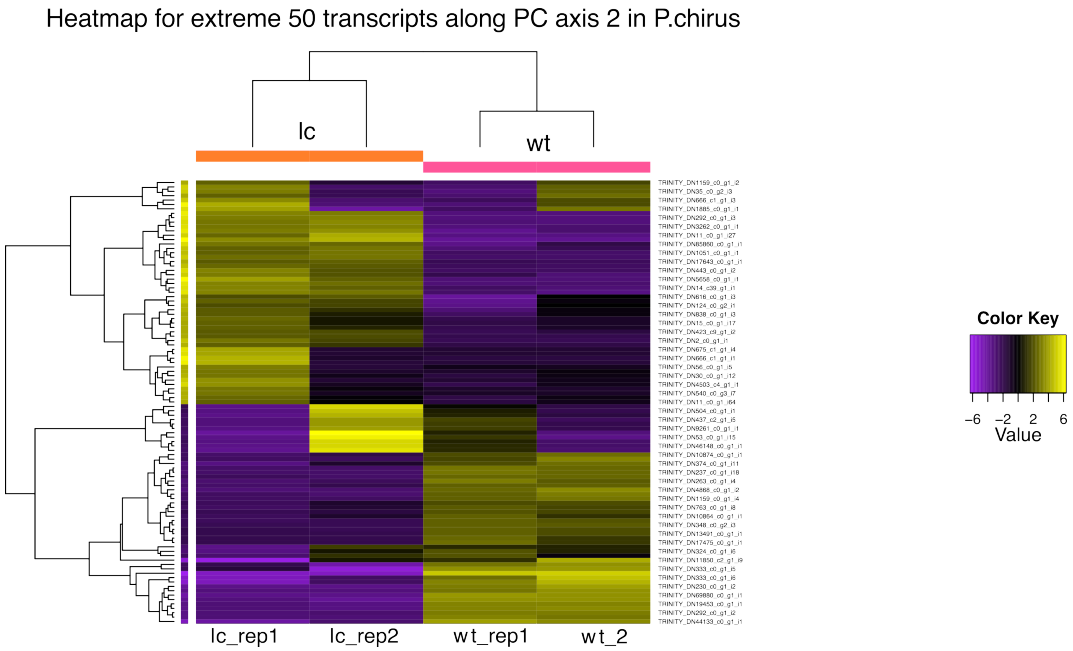

Figure 5: Subset heatmap for 50 extreme transcripts driving variation along the second axis of variation PC 2 for P.chiurs.