Differential Expression in *Hematodinium sp.*

Aidan Coyle

2021-11-09

```
library("kableExtra")
library("tidyverse")
library("knitr")
library("magick")
```

Methods

400 [TK: this was for the full experiment, should we reduce to, say, the 9 with libraries?] male *C. bairdi* were collected with pots from Stephen's Passage in southeastern Alaska in October 2017. Crabs were then transported to the Ted Stevens Marine Research Institute in Juneau, AK and placed in TK L flow-through seawater tanks. They were then (TK: assuming covered w/ insulating foam board) held at 7.5°C for a 9-day acclimation period. At the end of this acclimation period, 0.2 ml of hemolymph was drawn from each crab and preserved in 1200 µl RNAlater.

At the conclusion of the acclimation period, crabs were divided randomly into three treatment groups. The control group was held at 7.5°C. The water temperature of the other two groups (henceforth the elevated and decreased groups) was gradually changed to 10°C and 4°C, respectively. This change took place over two days. At the end of the two-day temperature adjustment, an additional 0.2 ml of hemolymph was drawn from each crab and preserved in 1200 µl RNAlater. Tanks were held at their temperatures for an additional 15 days, for a total experimental duration of 17 days. All surviving crabs then had three additional 0.2 ml hemolymph samples withdrawn and preserved in 1200 µl RNAlater. Due to a mass mortality event in the elevated group, no samples in this group were taken. Crabs were then humanely euthanized.

For samples from Day 17 of the experiment [TK: this looks like a real soft spot for any conclusions of high vs. low infection status for me - we're determining high vs. low after the majority of the samples were taken], DNA was extracted and subjected to qPCR following established protocol for *Hematodinium sp.* (Crosson 2011) and aligned to a provided species-specific standard curve. Samples were tested in duplicate. This provided a measure of the level of *Hematodinium sp.* infection.

RNA was extracted from all samples using Quick DNA/RNA Microprep Plus Kit (Zymo Research) according to the manufacturer's protocol [TK: Cite Zymo in references?]. Samples were quantified (2 µl) on Qubit 3.0 using the Qubit RNA HS Kit (Invitrogen) [TK: Cite Invitrogen in references?]. Based on RNA yield, three crabs were chosen from each treatment group, and all samples from these crabs were submitted to the Northwest Genomics Center at Foege Hall at the University of Washington for construction and sequencing of RNA-seq libraries.

Transcriptome Assembly and Annotation

Raw sequence data were assessed using FastQC (v0.11.8; Andrews 2010) and MultiQC (v1.6; Ewels et al. 2016) pre- and post-trimming. Data were quality trimmed using fastp (v0.20.0) (Chen et al. 2018). A transcriptome was *de novo* assembled from all libraries from the nine sequenced crabs, along with a number

of pooled samples (Supp. Table TK), using Trinity (v2.9.0; Grabherr et al. 2011; Haas et al. 2013). This is hereafter referred to as the unfiltered transcriptome. Trimmed sequencing reads were functionally annotated with DIAMOND BLASTx (0.9.26; Buchfink et al. 2015) using the UniProt Swiss-Prot database (downloaded 2020-01-23) [TK: Sprot = Swiss-Prot db right?].

To examine host expression specifically, the individual libraries used in the creation of the unfiltered transcriptome were compared and annotated using DIAMOND BLASTx (v0.9.29) to a publicly-available *Chionoecetes opilio* genome (NCBI Acc: GCA_016584305.1, citation TK). *C. opilio* and *C. bairdi* are quite closely related, and often produce viable hybrids. Sequences from the libraries with an e-value below 1x10ˆ-4 were kept and assembled using Trinity (v2.12.0) into a transcriptome. This is hereafter referred to as the host transcriptome.

A third transcriptome was created to examine expression in *Hematodinium sp.* The same set of libraries were imported into MEGAN6 (citation TK) and a taxonomic filter was applied to select only *Alveolata* sequences. These sequences were then assembled using Trinity (v2.9.0) and annotated with DIAMOND BLASTx (v0.9.29). This transcriptome is hereafter referred to as the parasite transcriptome

Library Alignment and Differential Expression Analysis

Every library was pseudoaligned to each of the three transcriptomes (unfiltered, host, and parasite) using kallisto (Bray et al. 2016), and abundance matrices were then produced using a perl script provided within the Trinity pipeline (v2.12.0). Pairwise comparisons for differential expression of contigs was performed with the R package DESeq2 (Love et al. 2014). Libraries were grouped based on treatment group, temperature at time of sample, and day for this series of pairwise comparisons. [TK: supplemental table of pairwise comparisons?]

Gene ontology (GO) terms were obtained by cross-referencing the accession IDs of each contig with the Gene Ontology database (TK citation). For each pairwise comparison, the log2-fold changes were extracted from the DESeq2 output. These were used as input for GO-MWU (Wright et al. 2015), which performs a Mann-Whitney U test and utilizes adaptive clustering to examine gene ontology term enrichment.

Characterizing Immune Genes

The cross-referenced table of accession IDs and GO terms for each transcriptome was filtered to examine genes with the GO term for "Immune Response" (GO: 0006955). Literature searches for functions within closely-related species were then performed

Network Analysis

The libraries produced by the pseudoalignments were used for three weighted correlation network analyses - one per transcriptome. For this, the R package WGCNA (Langfelder & Horvath 2008) was used. Contigs were clustered by expression patterns into module eigengenes. Those modules were then correlated with sample traits, such as crab, temperature, day, carapace width, and infection level as determined by qPCR.

Analyzing WGCNA Modules

All modules with a significant correlation to a sample trait were examined. If the significance appeared to be the result of correlation to libraries from a single crab, the module was discarded. The module membership (kME) of contigs belonging to that module was extracted and analyzed using GO-MWU.

Table 1: Individual libraries

Crab ID	Treatment group	Day 0 sample ID	Day 2 sample ID	Day 17 sample ID
A	Ambient	178	359	463
B	Ambient	118	349	481
C	Ambient	132	334	485
D	Decreased	73	221	427
E	Decreased	151	254	445
F	Decreased	113	222	425
G	Elevated	173	272	NA
H	Elevated	72	294	NA
I	Elevated	127	280	NA

TK section

Talk here about how lots of genes from the unfiltered transcriptome didn't match closely to either the host genome or parasite genes? Also did we make sure there was no overlap in transcripts between the two (i.e. none that were assigned to both the host and parasite transcriptome)?

Results

DESeq2

The DESeq2 package was used to examine differential expression between libraries, and to perform various pairwise comparisons between sample groups. Principal component analyses of samples taken from the elevated-temperature treatment group showed clustering by day, and thus by temperature. This was observed for libraries aligned to both the unfiltered and host-only transcriptomes. Due to low counts, a PCA could not be created for libraries aligned to the parasite-only transcriptome. No such clustering was observed for the ambient-temperature libraries, regardless of transcriptome, along this same timeframe.

[TK: CHANGE IMAGE LEGENDS FROM TEMPERATURE TO DAY. Done for hemat1.6, just rerun DESeq for other PCAs needed (after updating exp_design table)].

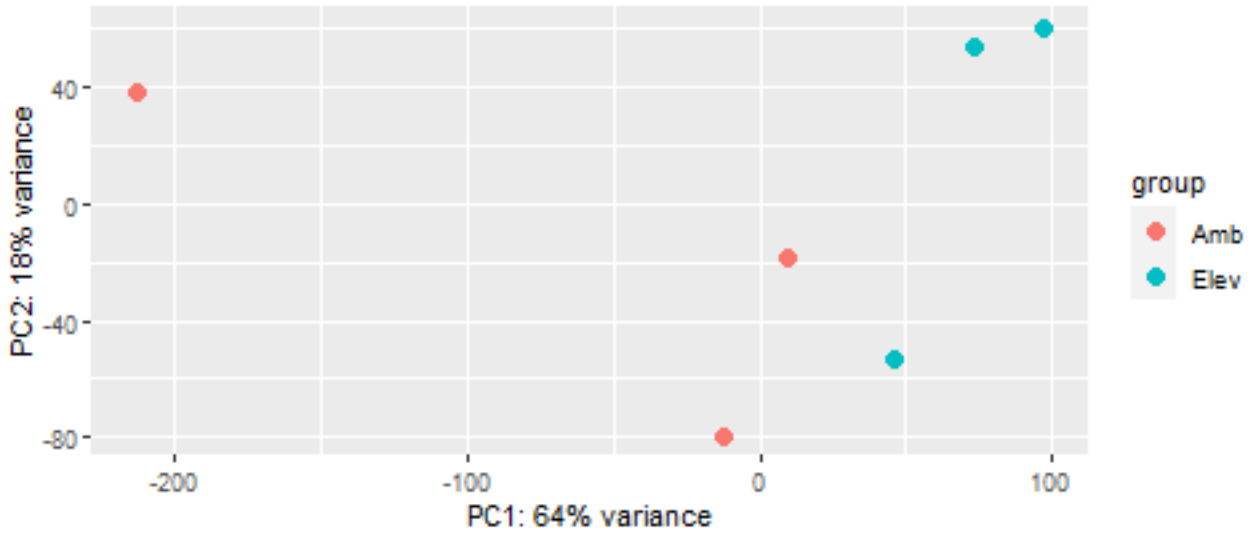

Figure 1: PCA for elevated-temperature libraries, Days 0-2 (unfiltered transcriptome)

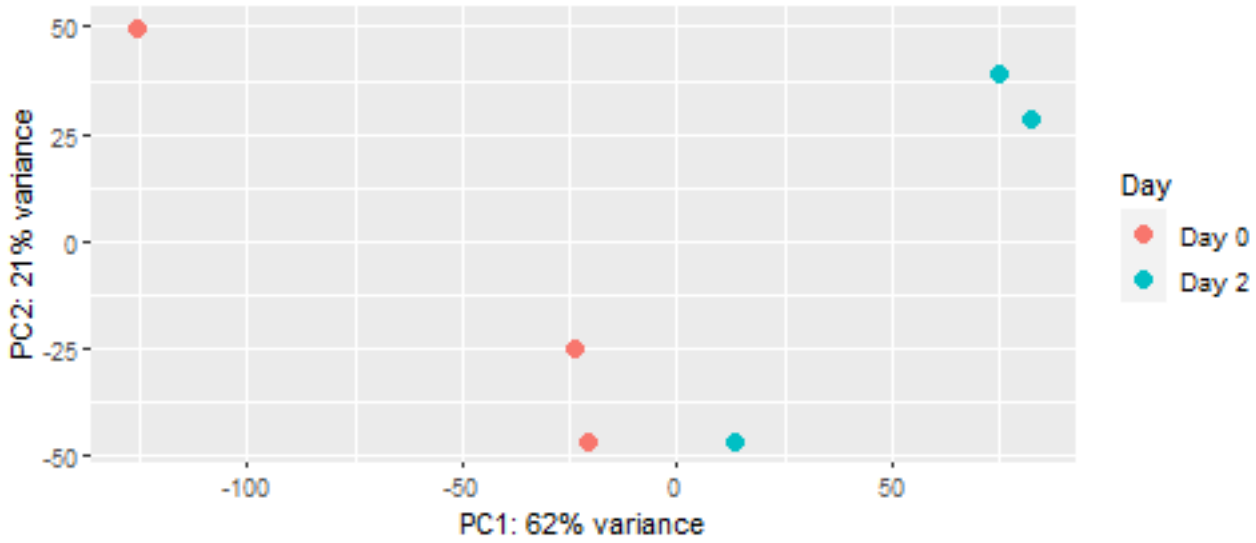

Figure 2: PCA for elevated-temperature libraries, Days 0-2 (crab transcriptome)

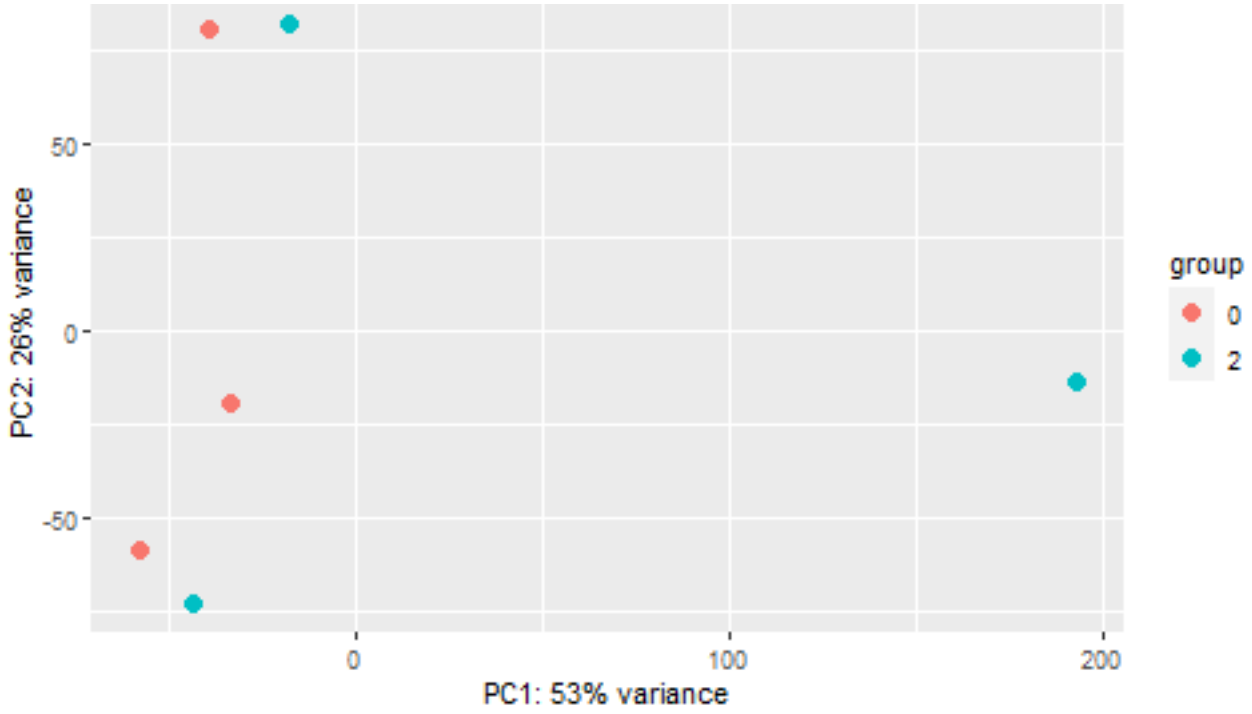

Figure 3: PCA for ambient-temperature libraries, Days 0-2 (unfiltered transcriptome)

Figure 4: PCA for ambient-temperature libraries, Days 0-2 (crab transcriptome)

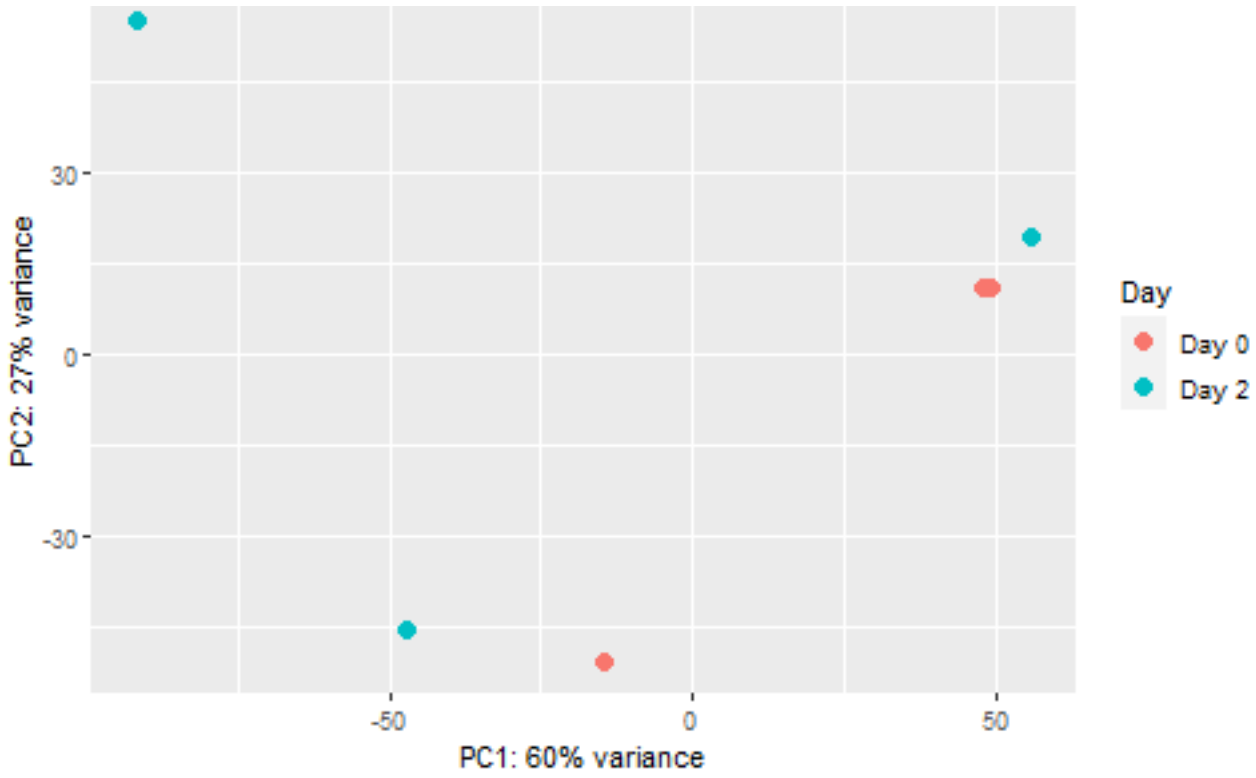

Figure 5: PCA for ambient-temperature libraries, Days 0-2 (parasite transcriptome)

GO-MWU

Pairwise comparisons were performed using GO-MWU to determine which biological processes were enriched.

Temperature Adaptation

Adaptation to temperature over the two-day temperature change period and, if applicable, the length of the experiment, was examined.

Unfiltered Transcriptome Unfiltered libraries from both the elevated-temperature treatment group saw enrichment in numerous biological processes over the two-day period of temperature change from ambient. Notably, enriched pathways included TK, TK, and TK. A large number of pathways were also enriched for the decreased-temperature treatment group over this same time period, including TK, TK, and TK. Within the ambient-temperature control group, process enrichment was minimal over this timespan.

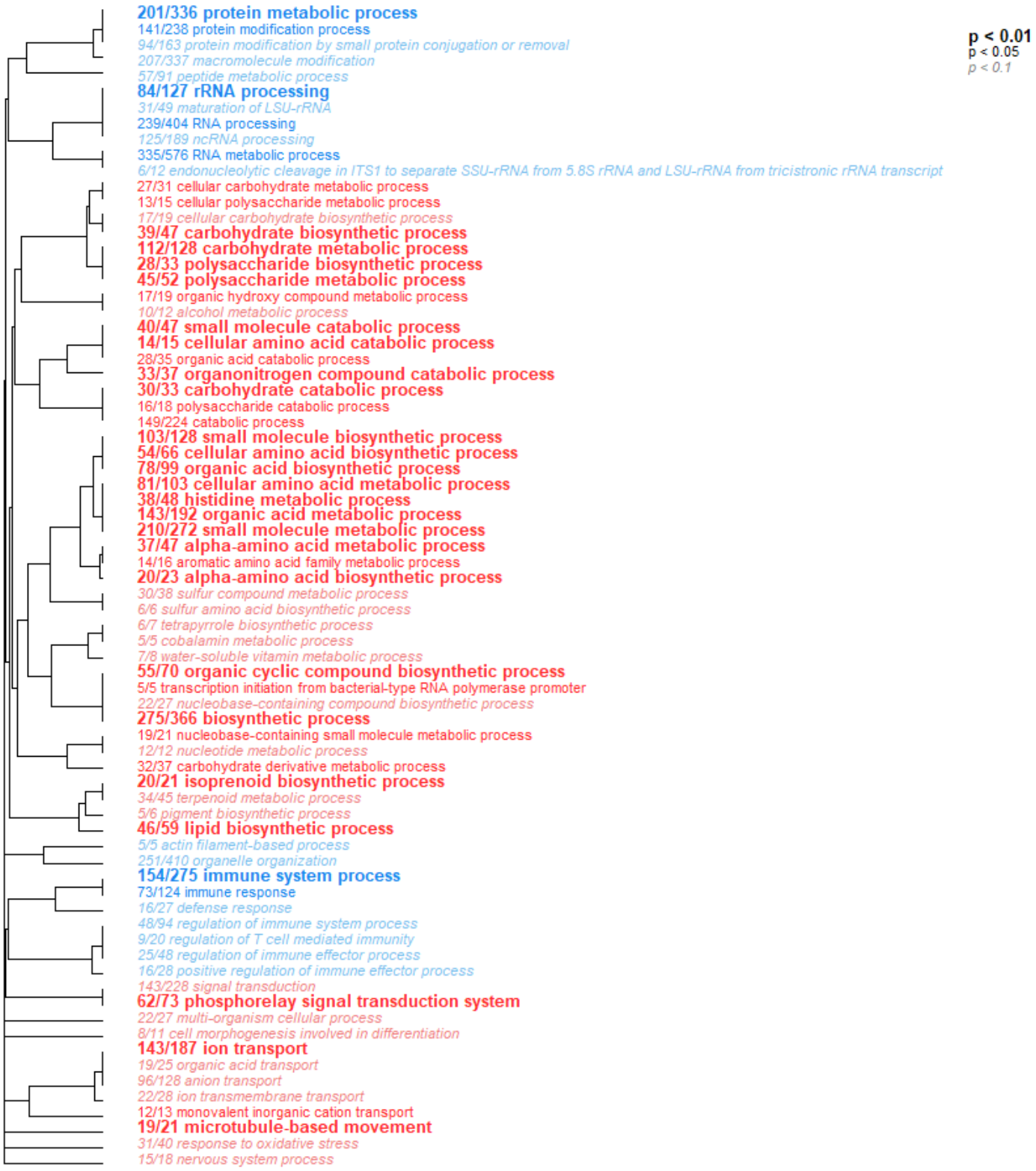

Figure 6: GO term enrichment for elevated-temperature libraries, Days 0-2 (unfiltered transcriptome)

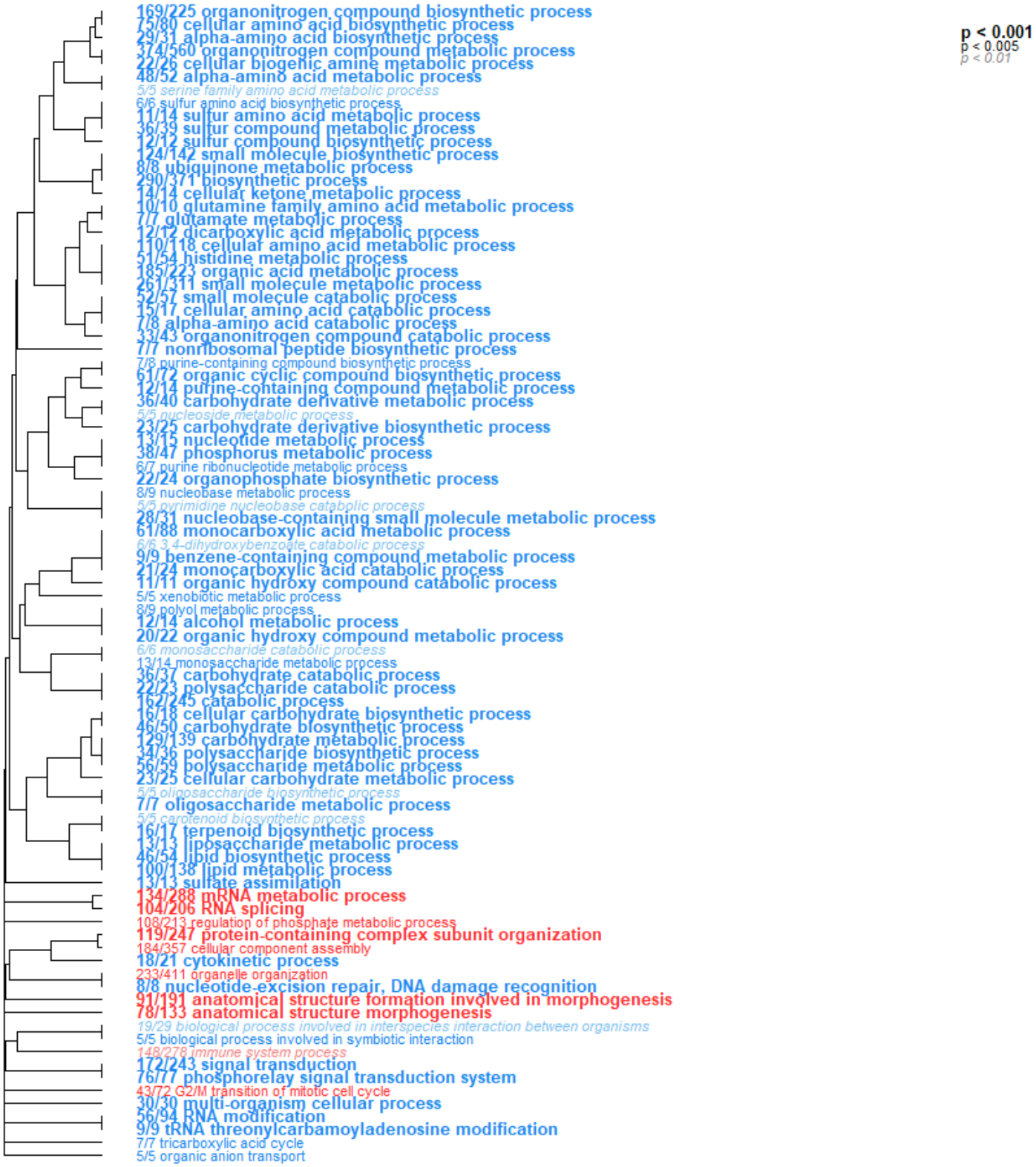

Figure 7: GO term enrichment for decreased-temperature libraries, Days 0-2 (unfiltered transcriptome)

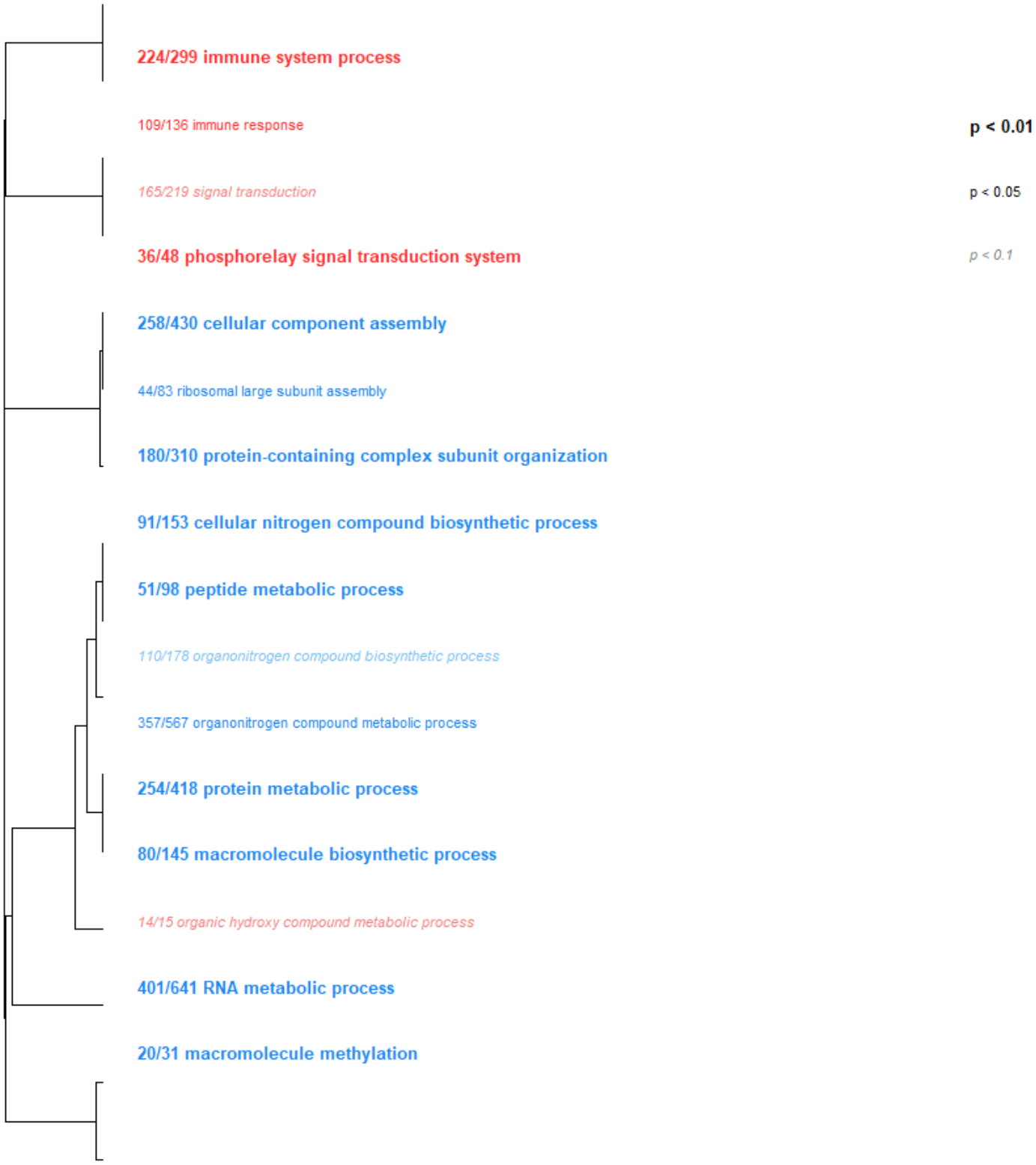

Figure 8: GO term enrichment for ambient-temperature libraries, Days 0-2 (unfiltered transcriptome)

Time

Unfiltered Transcriptome Over the 17 days of the experiment, GO term expression changed substantially within the control group. Overall, the control group saw changes in TK, TK, and TK pathways. Based on this, it appears that TK, TK, and TK pathways are involved in over the course of an infection with *Hematodinium*.

TK: DISCUSSION, TALK ABOUT HOW THIS COULD EITHER BE AN INDICATION OF CHANGES OVER COURSE OF INFECTION OR INDICATION OF TANK ADAPTATION.

Figure 9: GO term enrichment for ambient-temperature libraries, Days 0-17 (unfiltered transcriptome)

Host Transcriptome Over the course of the experiment, only minor pathway expression changes were observed. This indicates TK IMMUNOSUPPRESSIVE TALK HERE - IS OTHER STUFF GETTING IN AND CHANGING?

Parasite Transcriptome While no major changes were observed in the host, *Hematodinium* expression did shift substantially over the same time period. Generally, those changes were associated with TK and TK. This indicates that as *Hematodinium* multiplies within the host, TK DISCUSSION OF POSSIBLE CHANGES IN MORPHOLOGY AND EXPRESSION AS DENSITY INCREASES

Figure 10: GO term enrichment for ambient-temperature libraries, Days 0-17 (unfiltered transcriptome)

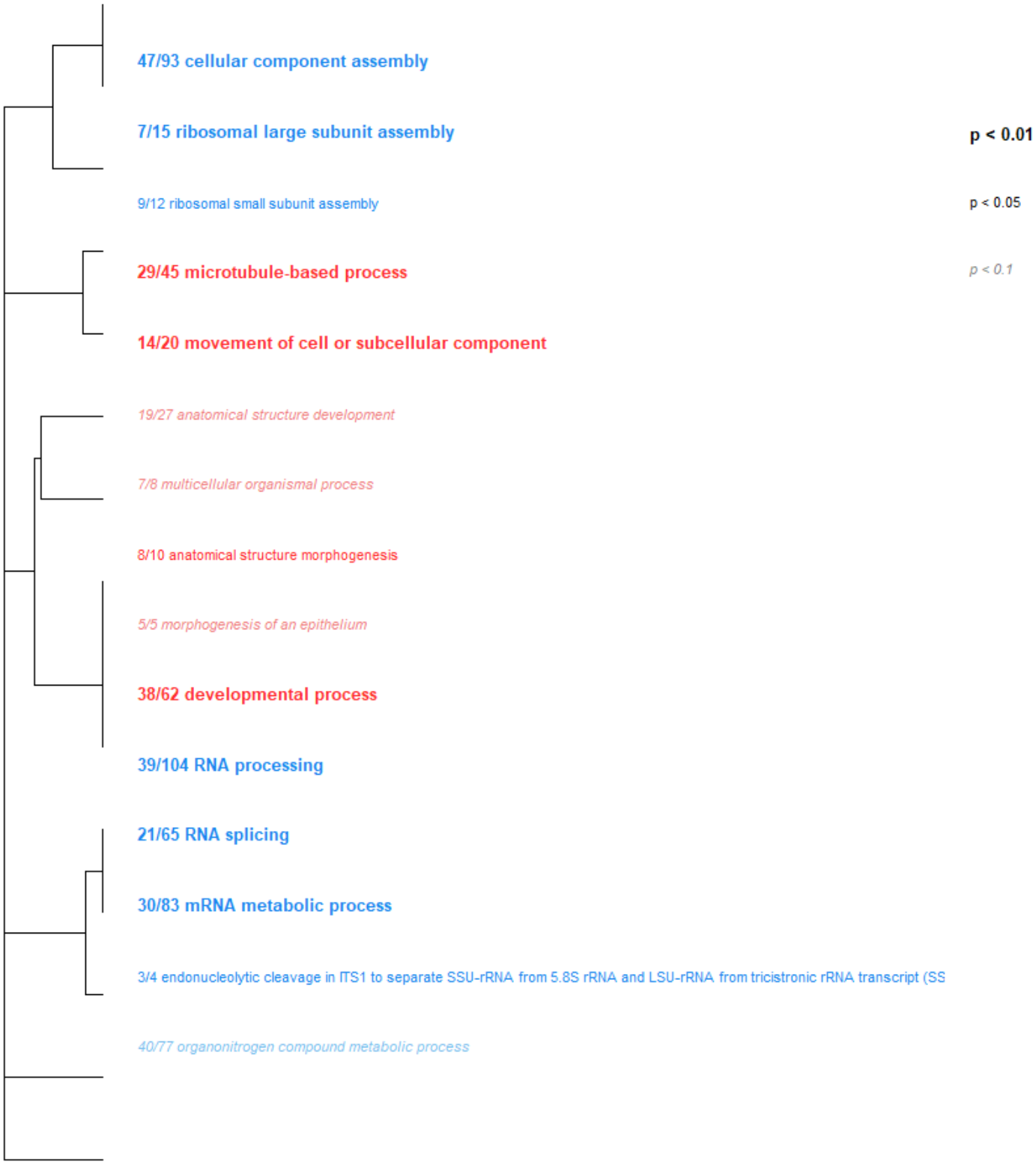

Figure 11: GO term enrichment for parasite, Days 0-17

Immune Genes

Host

Numerous genes (n = TK) within the *C. bairdi* transcriptome were associated with immune function (GO: 0006955). Many were members of the Cathepsin family, with Cathepsins C, J, L, S, U, V, and W all present. Cathepsin L was particularly broadly expressed, with seven distinct genes coding for Cathepsin and Procathepsin L [TK: does this make sense to say?]. Furthermore, Procathepsin L was differentially-expressed in the experimental group. Several types of MAPKs (mitogen-activated protein kinases) were also present within the transcriptome, including two p38 MAPKs and one one MAP4K. MAPKs are part of the IMD (immune deficiency) pathway, a notable component of the crustacean immune system. Several other genes associated with the IMD pathway were observed, including the transcription factor Relish and the kinase inhibitor IκK [TK: change 1st K to a kappa]. NFIL3, a nuclear factor which has been found to regulate Relish expression in similar systems, was also present.

Other notable immune-linked genes observed were Transcription Activator Protein-1 (TF AP-1) and Granzyme A. TF AP-1 acts as an immune system regulator within other crab species, along with a potential role as an osmoregulator. Little research on the role of Granzyme A in invertebrates has been performed, but in vertebrates it has a cytotoxic role against intracellular pathogens.

Parasite

Within the *Hematodinium sp.* transcriptome, 4 genes were linked to immune function. All four of these were cysteine proteases, which TK CP DESCRIPTION. Three of the four were cathepsins, including both Procathepsin and Cathepsin L. TK: PROBBALY TALK ABOUT ROLE OF CAT L IN PARASITES WITHIN DISCUSSION, BUT NEED A BIT MORE TO ROUND THIS SECTION OUT

TK: SOME SORT OF TABLE FOR DESCRIBING IMMUNE GENES?

Characterizing Overall Expression Patterns

Prior to filtering by taxa, samples from the lowered-temperature treatment group saw an average overall decrease in expression in 42% of transcripts, while the control group averaged a 33% decrease

Table TK: Overall expression in samples unfiltered by taxa

	Ambient	Lowered
Increase	30.8%	27.4%
Decrease	33.9%	42.9%
Neither	35.5%	29.6%

However, this same pattern was not observed when examining expression within the host or parasite specifically. Within the host, overall expression patterns were remarkably similar regardless of temperature. And within the parasite, expression increased within the lowered-temperature treatment group (TK STATISTICAL TESTS ON THESE RESULTS - CHI-SQUARE?) TK: MENTION CAVEAT OF 2 UNINFECTED IN LOWERED-TEMP GROUP, OR DO IN DISCUSSION?

Table TK: Overall host transcript expression

	Ambient	Lowered
Increase	28.9%	28.6%
Decrease	31.8%	32.2%

	Ambient	Lowered
Neither	39.3%	39.2%

Table TK: Overall parasite transcript expression

	Ambient	Lowered
Increase	32.3%	43.4%
Decrease	29.5%	29.1%
Neither	42.3%	30.1%

Characterizing Immune Gene Expression Patterns

NOTE: I really don't think our sample size of immune genes is large enough to make overall judgments on expression patterns, so this section is probably ripe for cutting. Still, I'll wait till I run a chi-square (or similar) test on this to see.

I'll also avoid writing up a more detailed analysis until I run those tests

Table TK: Immune gene expression in samples unfiltered by taxa

	Ambient	Lowered
Increase	37.1%	8.1%
Decrease	28.5%	67.2%
Neither	34.4%	24.7%

Table TK: Immune gene host transcript expression

	Ambient	Lowered
Increase	31.4%	9.7%
Decrease	28.5%	57.6%
Neither	40.0%	32.7%

Parasite expression: not available, only 5 genes total

WGCNA

A signed weighted correlation network analysis (WGCNA) was run on all libraries aligned to each transcriptome (TK citation). This clustered genes into modules according to expression pattern, and then correlated them with our variables. We took all modules that were significantly correlated, and discarded those in which the correlation to the variable appeared to be due to extremely strong correlation to a single crab. This produced the following modules (Table TK).

Transcriptome	Module	Trait and p-values
Unfiltered	black	Day (0.04)
Unfiltered	tan	Low vs. Ambient (0.05)
Unfiltered	cyan	Low vs. Ambient (0.02), Elevated vs. All (0.04)
Unfiltered	brown	Low vs. Elevated (0.02), Elevated vs. All (0.03)

Transcriptome	Module	Trait and p-values
Host	black	Low vs. Ambient (6x10^-4), Elevated (0.05)
Host	red	Low vs. Ambient (0.01), Infection Level (0.01), Carapace Width (0.04)
Host	blue	Low vs. Elevated (0.02), Elevated (0.03)
Parasite	black	Day (0.04)
Parasite	turquoise	Low vs. Ambient (0.02), Infection Level (0.01)
Parasite	blue	Infection Level (0.01)

Each of these modules was then analyzed using GO-MWU. No groups of GO terms were differentially enriched for any unfiltered or host module. However, all three modules within the parasite transcriptome saw differential enrichment. TK: Discussion of differential enrichment comes here

TK: Consider running DESeq2 on Hemat_Level H vs L

TK: Also consider running DESeq with contrasts to run three-way comparison on Amb 0 vs 2 vs 17 all in one

TK: Check whether, for the All Crabs PCAs, I should put multiple legends into the plot

TK: When we decide when/if to use WGCNA heatmaps, expand em to make em prettier

Literature Cited

Andrews S. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data". Available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bray, N.L., Pimentel, H., Melsted, P., & Pachter, L. 2016. "Near-optimal probabilistic RNA-seq quantification". *Nature Biotechnology* 34: 525-527 doi:10.1038/nbt.3519

Buchfink, B., Xie, C., & Huson, D.H. 2015. "Fast and sensitive protein alignment using DIAMOND". *Nature Methods* 12(1): 59-60. doi: 10.1038/nmeth.3176

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

Crosson, L.M. 2011. "Development and validation of a quantitative real-time polymerase chain reaction (qPCR) assay to assess the impact of Hematodinium, a parasitic dinoflagellate, on Tanner crab populations in Alaska". University of Washington

Ewels, P., Magnusson, M., Lundin S., & Käller M. 2016. "MultiQC: summarize analysis results for multiple tools and samples in a single report". *Bioinformatics* 32(19): 3047-3048 doi:10.1093/bioinformatics/btw354

Grabherr, M.G., Hass, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X. et al. 2011. "Full-length transcriptome assembly from RNA-seq data without a reference genome". *Nature Biotechnology* 29: 644-652 doi:10.1038/nbt.1883

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Cougar, M.B. et al. 2013. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis". *Nature Protocols* 8(8): 1494-1512 doi:10.1038/nprot.2013.084

Langfelder, P., Horvath, S. 2008. "WGCNA: an R package for weighted correlation network analysis". *BMC Bioinformatics* 9: 559 doi:10.1186/1471-2105-9-559 doi:10.1186/1471-2105-9-559

Love, M.I., Huber, W., Anders, S. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15: 550. doi:10.1186/s13059-014-0550-8

Wright, R.M., Aglyamova, G.V., Meyer E., & Matz M.V. 2015. "Gene expression associated with white syndromes in a reef building coral, *Acropora hyacinthus*". *BMC Genomics* 16: 371. doi:10.1186/s12864-015-1540-2

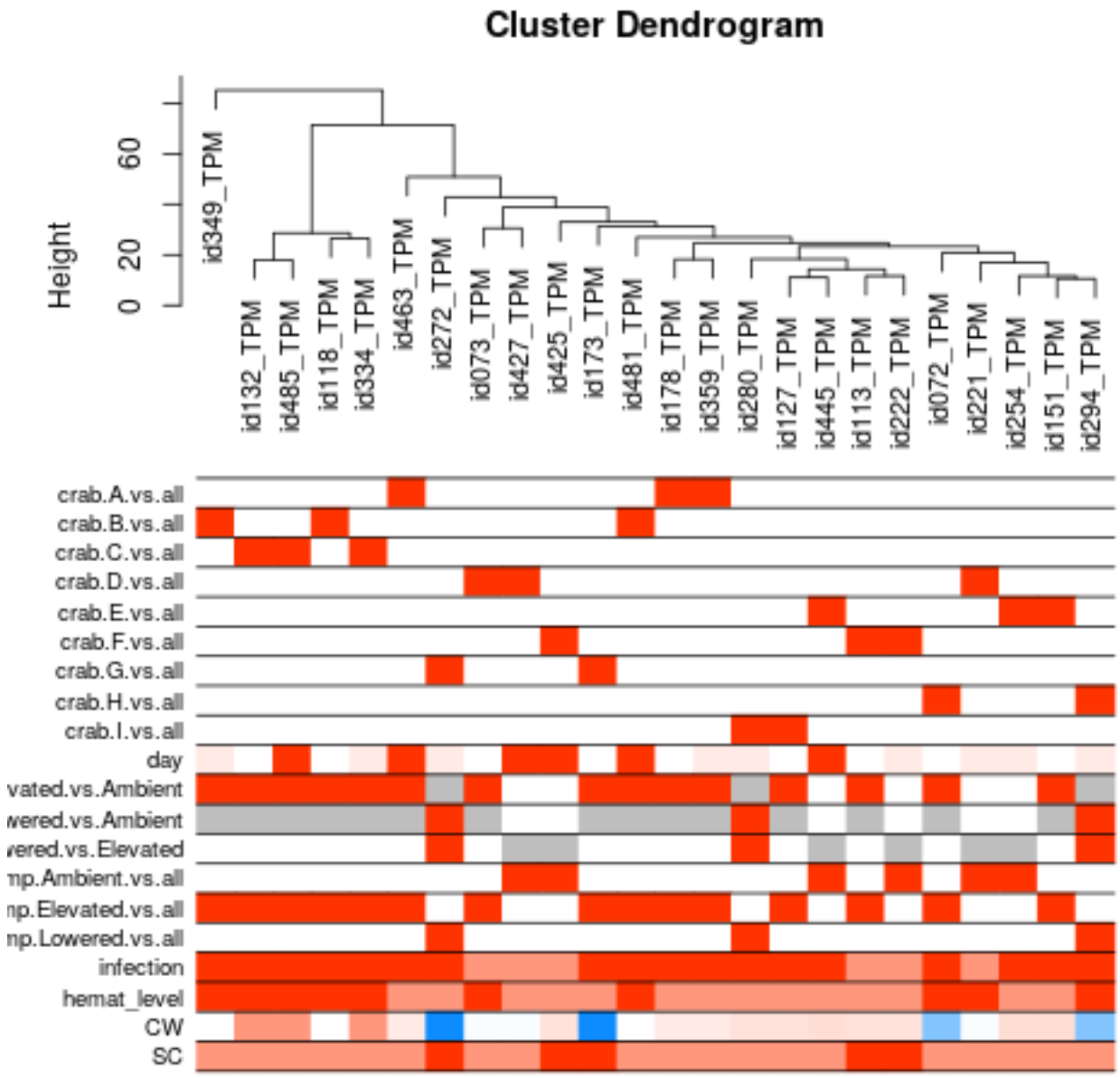

Figure 12: WGCNA Cluster Dendrogram for traits in parasite libraries

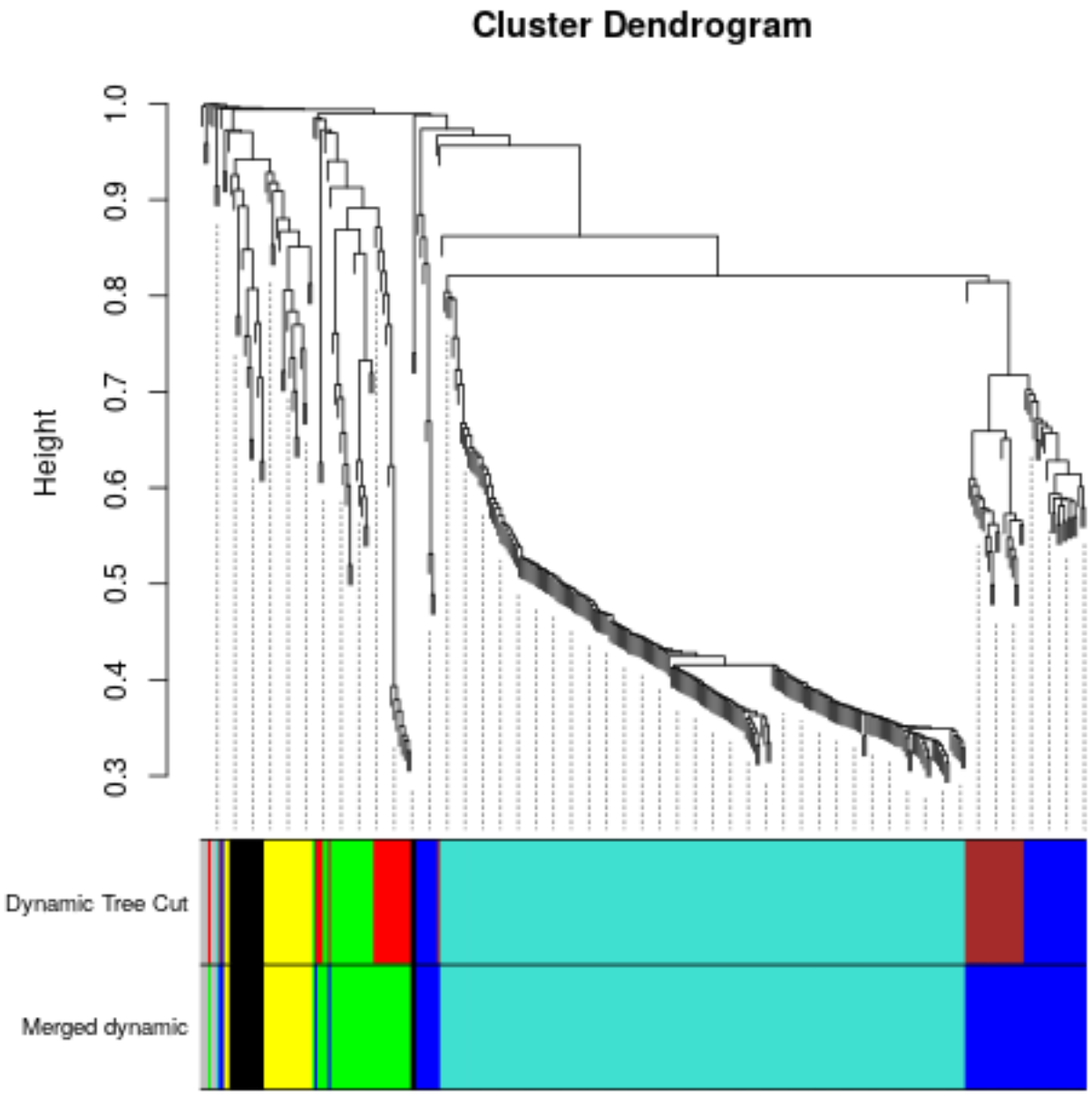

Figure 13: WGCNA Cluster Dendrogram of original and merged eigengenes for parasite libraries

Figure 14: GO term enrichment for parasite black module (linked to day)

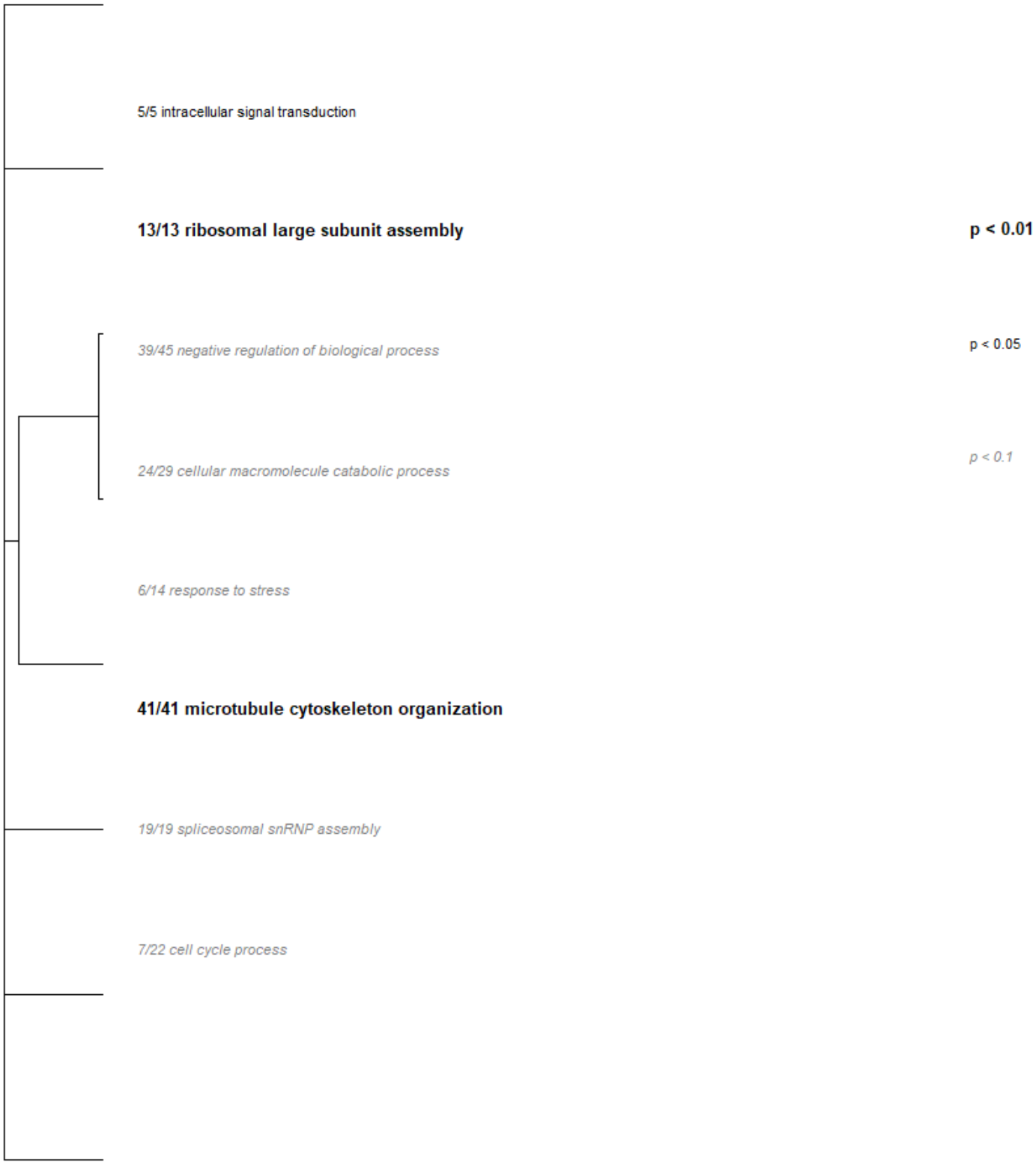

Figure 15: GO term enrichment for parasite turquoise module (linked to temperature and infection level)

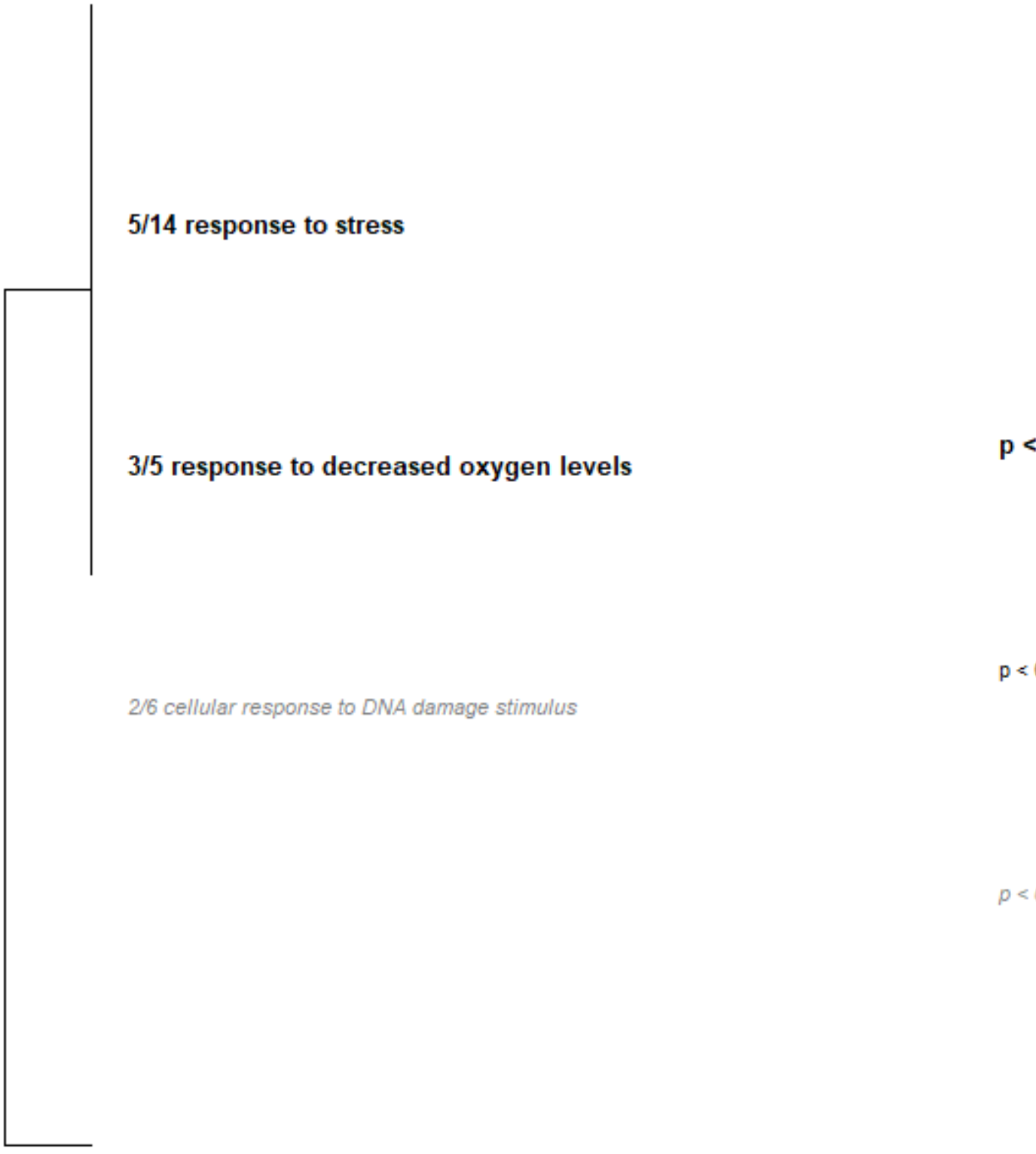

Figure 16: GO term enrichment for parasite blue module (linked to infection level)