# Red King Crab ‘*Paralithodes camtschaticus*’ RNA-Seq Project *continued*

Laura Spencer’s notes on bioinformatics and data analysis

**Important file locations:**

* GitHub repo - <https://github.com/laurahspencer/red-king_RNASeq-2022>
* Jupyter notebook for data processing - [red-king\_RNASeq-2022](https://nbviewer.org/github/laurahspencer/red-king_RNASeq-2022/blob/main/notebooks/Notebook%2001%20-%20read%20quality%20and%20trimming.ipynb)
* RMarkdown notebook with DESeq2 analysis: [02-DESeq2-etc-RKC.html](https://htmlpreview.github.io/?https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/notebooks/02-DESeq2-etc-RKC.html)

**Raw data location:**

* Raw sequence data concatenated by tank+crab is backed up to Google Drive ([RKC\_RNAseq/raw-data](https://drive.google.com/drive/folders/1Olk2fKT8-YEbYSgDpdaO1QlBE66qTQpb?usp=sharing)) and is located on Sedna in /share/nwfsc/ggoetz/red\_king\_crab/illumina/

**Draft manuscript**: [RKC OA RNASeq Manuscript Draft](https://docs.google.com/document/d/14SimrkmnbMuzcnre0v0n78uJtoP_AuwKzvKc13o02aM/edit?usp=sharing)

Broad Task: Describe gene expression patterns in gill tissue of Red king crab reared in varying pH conditions. Which genes and processes shift when individuals are reared/acclimated to varying pH environments?

Overview of pipeline:

* Concatenate data from different lanes by Tank and Crab number (Giles already did this, location of concatenated files on Sedna: share/nwfsc/ggoetz/red\_king\_crab/illumina/)/
* Assess quality via FastQC / MultiQC
* Clean raw RNASeq data - remove adapters, poly-A tails, removal of N’s, filter for minimum size (Giles used 50bp) and perhaps quality(?) using Cutadapt.
  + NOTE: Giles included his cutadapt script in the 0011 file
* Reassess quality and confirm adapter removal via FastQC / MultiQC
* Align reads to [blue king crab genome](https://www.ncbi.nlm.nih.gov/genome/?term=txid273748%5Borgn%5D) and generate gene-level counts using:
  + Option 1: using Bowtie2 and [featureCounts OR RSEM]
  + Option 2: using STAR
  + NOTE: also would like to generate isoform-level counts. Figure out which program does that.
* OPTIONAL STEP: Deduplicate aligned reads using Picard (FastQC could inform, also could run analysis w/ and w/o deduplicating to see how it affects things)
* Perform differential expression analysis using DESeq2 (count normalization not necessary)
* Perform cluster analysis - either using DESeq2 or my own code for PCA, perhaps another multivariate analysis such as partial least squares-discriminant analysis (PLS-DA, what Ariana prefers)
* Identify clusters (modules) of highly correlated genes by pH (and other factors) using WGCNA (may need to normalize counts - figure out)
* Perform pathway analysis (IPA)
* OPTIONAL ANALYSIS: Characterize relatedness among individuals by calling SNPs and clustering (?)

## December 8 & 9, 2021

Gained access to Sedna and red king crab data. Got Cygwin, git, jupyter notebook, atom installed on new computer. Explored blue king crab genome on NCBI and publication. Began literature review of red king crab OA research to date - see Paperpile folder.

ISSUE TO BE RESOLVED: Blue king crab genome on NCBI doesn’t include annotation files. I emailed the genome [publication](https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13266) authors and NCBI to see if annotation files are published elsewhere. Waiting to hear back.

## Dec 13, 2021

Got familiar with data, sketched out a pipeline. Created a github repo for this project, <https://github.com/laurahspencer/red-king_RNASeq-2022> and started a Jupyter notebook to document my work/code - [red-king\_RNASeq-2022](https://nbviewer.org/github/laurahspencer/red-king_RNASeq-2022/blob/main/notebooks/Notebook%2001%20-%20read%20quality%20and%20trimming.ipynb). Ran my first job on Sedna - fastqc and multiqc on the raw (but concatenated) sequence files.

## Dec 14, 2021

Transferred MultiQC report of raw sequence files from Sedna to local computer. Created slurm script based off of Giles’ code to trim and filter reads using cutadapt, then run fastqc and multqc on the trimmed data. Started job around 6:30pm.

## Dec 16, 2021

Re-trimming with slightly more stringent quality settings (phred=20 on 5’ and 15 on 3’, compared to 15 and 10, respectively). Heard back from NCBI - no annotation files were submitted with the blue king crab genome. Re-emailing authors to follow-up.

## Dec 27, 2021

Finally got Bowtie2 to start working on Sedna. Started building the index for the blue king crab genome and aligning reads.

## Dec 30, 2021

Bowtie2 finished aligning trimmed reads to blue crab genome. Output file is here: [bowtieout.v1.txt](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/results/bowtie/bowtieout.v1.txt). Alignment rates look pretty good:

* Average alignment of all samples = 78.44% +/- 1.9 (SD)
* Alignment ranged from 73.77% to 81.26%

Started converting alignment .sam files to .bam files, sort by coordinate, then deduplicate using gatk MarkDuplicates.

## Jan 2, 2022

File conversion, sort & deduplicate complete. Resulting .bam files are located on Sedna here: /scratch/lspencer/2022-redking-OA/aligned-bowtie/sorted-deduped and include three files per sample:

* Tank\_10\_Crab\_1.sorted.bam (alignement data prior to deduplicating)
* Tank\_10\_Crab\_1.dedup.bam
* Tank\_10\_Crab\_1.dup\_metrics.txt

The average duplication rate was 31.3%, and ranged from 24.3%-59.3% (pretty high). I won’t likely use the deduplicated data, but will need to explore the non-deduplicated data in IGV to confirm that this is the right choice.

## Jan 3, 2022

Good news! I received two blue king crab genome annotation files from the authors of [Tang et al. 2020](https://doi.org/10.1111/1755-0998.13266) with locations of various features (EVM.out\_new.gff3.gz) and just coding sequences (EVM.out\_new.cds.gz). They are saved on Sedna here: /home/lspencer/references/bluekingcrab/ and on GitHub here: <https://github.com/laurahspencer/red-king_RNASeq-2022/tree/main/references>

Hiccup: the annotation files use the blue king crab genome version published on [Dryad](https://datadryad.org/stash/dataset/doi:10.5061/dryad.jm63xsj6c), but I used the version published on NCBI when aligning. They have different chromosome & scaffold IDs. I will see if I can translate between the two, but worst case scenario I’ll re-align to the Dryad version (probably safest).

Started re-aligning trimmed RNASeq data (trimmed version 2, which used slightly more conservative quality filtering settings) to the Dryad version of the blue king crab genome using the slurm script [2022-01-03\_align.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/2022-01-03_align.sh)

## Jan 4, 2022

Started exploring the blue king crab feature files. Have some questions about nomenclature & number of scaffolds listed in .fasta and what they represent. I emailed authors of Tang et al. 2020 with question:

*I understand that you assembled a chromosome-level genome, with 104 chromosomes and 6,958 scaffolds. The genome .fasta posted to Dryad includes scaffolds numbered from "HiC\_scaffold\_1" up to "HiC\_scaffold\_6958". However, I see that scaffolds 1-104 are nearly the lengths of the chromosomes listed in Table 3 of your publication, so* ***do scaffolds #1-104 in the Dryad genome represent the chromosomes?*** *If so, are scaffolds 105 to 6,958 additional sequences (i*.e. unplaced scaffolds), or are they sub-sequences of the chromosomes (i.e. placed scaffolds)?

Seems like a silly question, however I would presume that, if the .fasta includes both the chromosomes AND the scaffolds (which are related hierarchically), there should be a total of 7,062 sequences (the 104 chromosomes + the 6,958 scaffolds)

## Jan 5, 2022

Alignment (version 2, to Dryad) finished. Resulting files are located in /home/lspencer/2022-redking-OA/scripts and labeled (for example) Tank\_10\_Crab\_1.bowtie.v2.sam. Started sorting and converting to .bam, then deduplicating using the script [2022-01-04\_sort-deduplicate.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/2022-01-04_sort-deduplicate.sh).

Alignment rates were the same as with the NCBI version of the genome, and averaged 78.44%.

Next step- generate gene counts using the .gff feature file. Submitted request for Subread (package that contains featureCounts) to be installed on Sedna.

## Jan 7, 2022

FeatureCounts finished without error

Started importing and reviewing gene counts in RStudio.

[1] "Number of samples: 43"

[1] "Total number of genes in dataframe: 28,287"

[1] "Average number of genes per sample: 20,138.72"

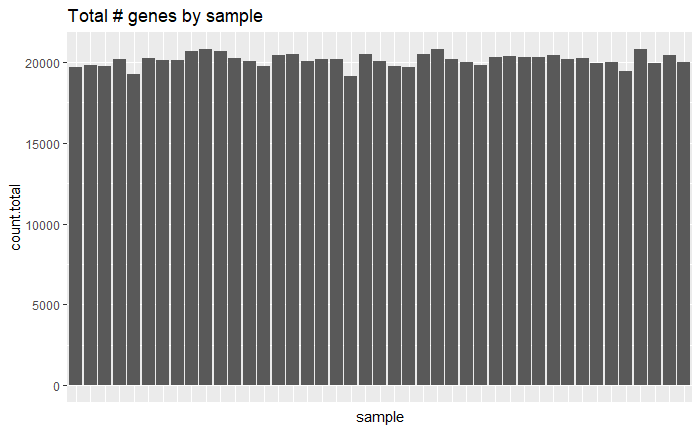
[1] "Total counts, all samples: 423,387,193"

Issue- the feature file identifies the locations of genes, BUT it does not describe/annotate them. I will need to blast the .cds feature file to identify gene function.

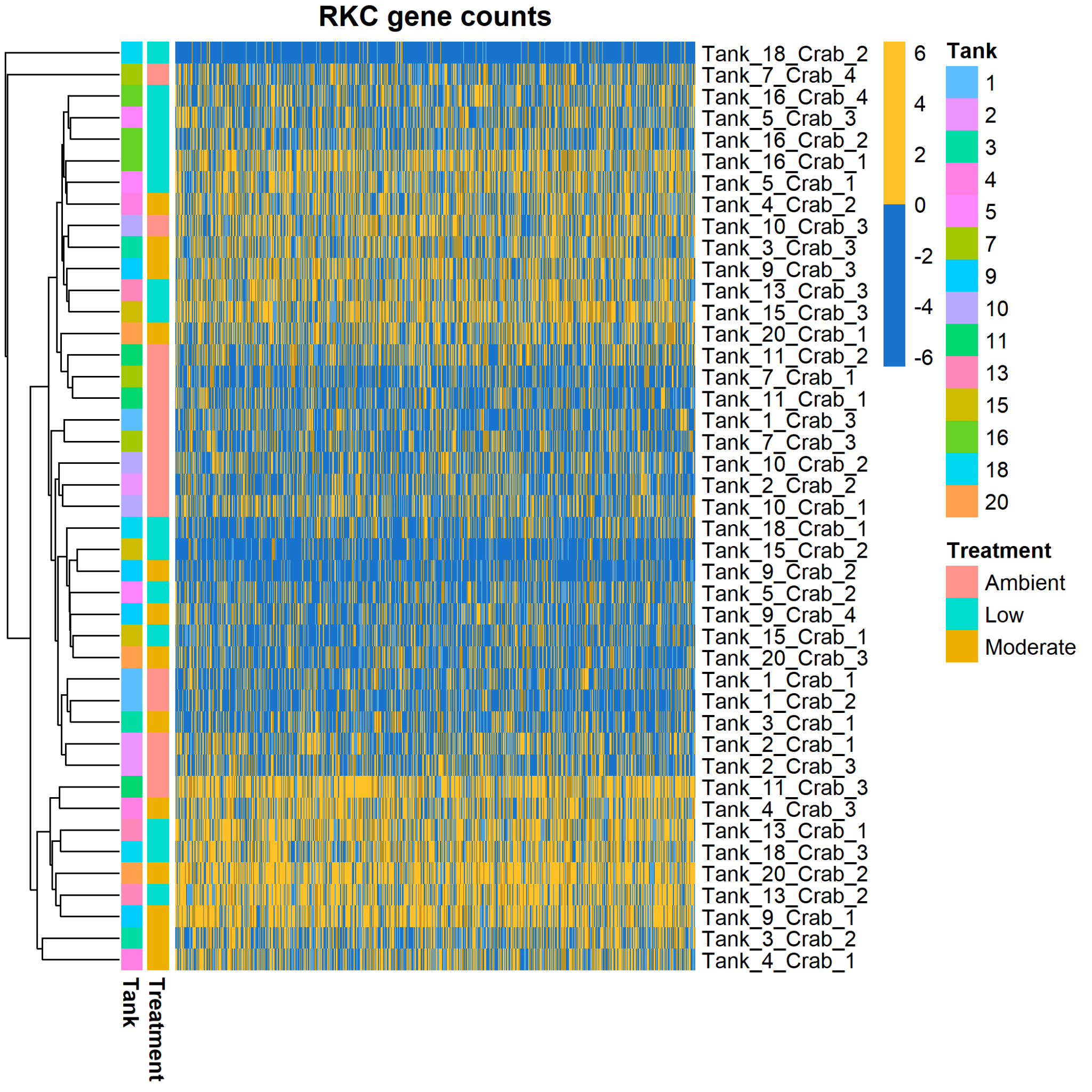
## Jan 9, 2022

Started running count data through DESeq2. Preliminary results:

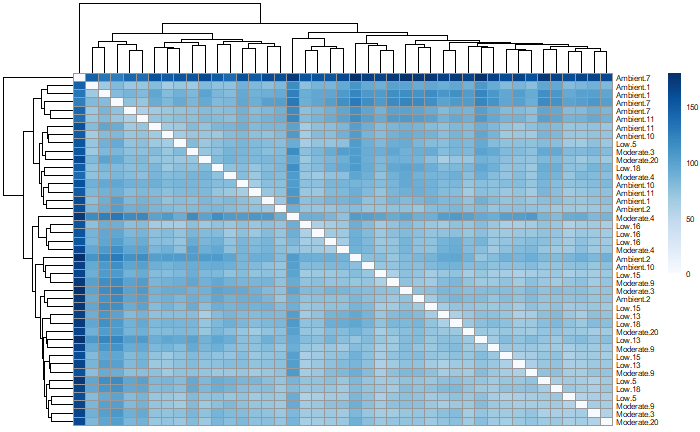
* Average # genes detected per sample: 20,139 +/- 381 (SD)
* # genes per sample ranges from 19,114 - 20,809

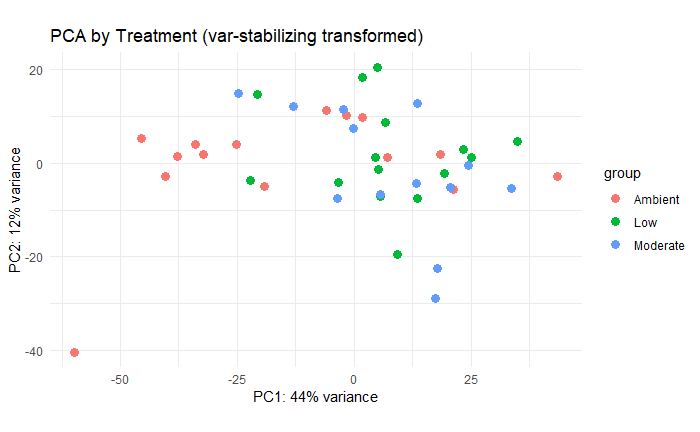


**Heatmap of all gene counts, clustered**



**Heatmap of sample-to-sample distances**





**Number of differentially expressed genes among the pH treatments:**

* Ambient pH vs. Moderate pH (7.8) = 1,412 genes
* Ambient pH vs. Low pH (7.5) = 2,097 genes
* Moderate pH (7.8) vs. Low pH (7.5) = 15 genes

## Jan 11, 2022

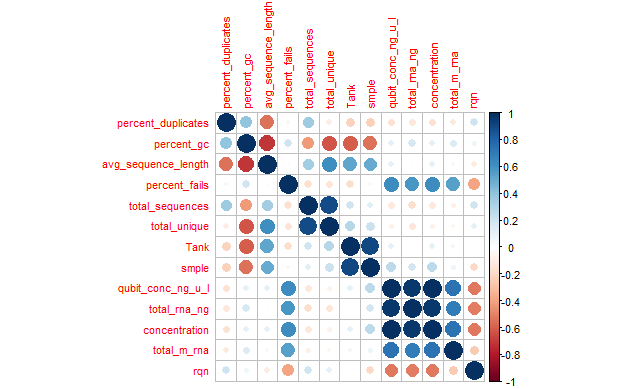
Started annotating the blue king crab coding sequence .fasta file with the UniProtKB/Swiss-Prot database using blastx.

## Jan 12, 2022

Realized it would be easier if blasted nucleotide sequences for genes instead of coding sequences. Generated a gene.fasta for the blue king crab, then started blast on that. Realizing now that I probably won’t get per-gene blast results from Giles’ run, though. Damn.

Tested out removing counts for genes mapping to “scaffolds” 105+, since I don’t know what those are (are they included in the chromosomes 1-104?). Also tested additional featureCounts settings. Need to start spreadsheet with resulting DEG numbers (but those changes didn’t have any dramatic effects to the PCAs or heat maps).

Obtained and explored library prep data. File I imported into R is available here: [library-prep-stats.csv](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/data/library-prep-stats.csv). Generated correlation plot of sequencing & library prep stats. Interesting correlations are between percent\_fails and library prep stats (e.g. percent\_fails ~ RQN). Good to see that percent\_duplicates doesn’t correlate with anything. All libraries were amplified using 12 pcr cycles, so that isn’t included here.



## Jan 14, 2022

My interim blast finished, and I merged the results with the gene count matrix, available on GitHub here: [results/counts.annotated.csv](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/results/counts.annotated.csv).

Did a very quick gene enrichment analysis using [DAVID](https://david.ncifcrf.gov/tools.jsp). There are 22 enriched biological processes (listed here: [DAVID-Enriched-BP\_Amb-Mod.txt](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/DAVID-Enriched-BP_Amb-Mod.txt) and below) in the 1,648 DEGs identified between pH 8.0 and 7.8, and 23 enriched biological processes (listed here: [DAVID-Enriched-BP\_Amb-Low.txt](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/DAVID-Enriched-BP_Amb-Low.txt) and below) in the 2,181 DEGs identified between pH 8.0 and pH 7.5. Some usual suspects!

**Ambient vs Moderate comparison, listed from most to least sign.:** *Now updated with results after removing outlier sample*

* GO:0006412~translation (14 genes) (7 w/o outlier)
* GO:0051603~proteolysis involved in cellular protein catabolic process (6 genes) (3 w/o outlier)
* ~~GO:0045454~cell redox homeostasis (6 genes) (not sign. w/o outlier)~~
* GO:0015991~ATP hydrolysis coupled proton transport (5 genes) (4 w/o outlier)
* ~~GO:0034599~cellular response to oxidative stress (5 genes) (not sign. w/o outlier)~~
* GO:0006379~mRNA cleavage (4 genes) (3 w/o outlier)
* ~~GO:0055114~oxidation-reduction process (16 genes) (not sign. w/o outlier)~~
* ~~GO:0006270~DNA replication initiation (5 genes) (not sign. w/o outlier)~~
* GO:0002181~cytoplasmic translation (4 genes) (3 w/o outlier)
* ~~GO:0006891~intra-Golgi vesicle-mediated transport (4 genes) (not sign. w/o outlier)~~
* GO:0046034~ATP metabolic process (4 genes) (3 w/o outlier)
* ~~GO:0021510~spinal cord development (4 genes) (not sign. w/o outlier)~~
* ~~GO:0007568~aging (4 genes) (not sign. w/o outlier)~~
* ~~GO:0006364~rRNA processing (6 genes) (not sign. w/o outlier)~~
* ~~GO:0045944~positive regulation of transcription from RNA polymerase II promoter (20 genes) (not sign. w/o outlier)~~
* ~~GO:0007400~neuroblast fate determination (3 genes)~~
* ~~GO:0015992~proton transport (3 genes) (not sign. w/o outlier)~~
* ~~GO:0006749~glutathione metabolic process (3 genes) (not sign. w/o outlier)~~
* ~~GO:0006890~retrograde vesicle-mediated transport, Golgi to ER (4 genes) (not sign. w/o outlier)~~
* GO:0010976~positive regulation of neuron projection development (4 genes) (3 w/o outlier)
* ~~GO:0006457~protein folding (7 genes) (not sign. w/o outlier)~~
* ~~GO:0006099~tricarboxylic acid cycle (5 genes) (not sign. w/o outlier)~~
* GO:0042493~response to drug (5 genes, w/o outlier)
* GO:0006379~mRNA cleavage (4 genes, w/o outlier)
* GO:0050770~regulation of axonogenesis (3 genes, w/o outlier)
* GO:0007474~imaginal disc-derived wing vein specification (3 genes, w/o outlier)

**Ambient vs Low comparison, listed from most to least sign.:**

* GO:0006412~translation (18 genes) (16 w/o outlier)
* ~~GO:0008152~metabolic process (16 genes) (not sign. w/o outlier)~~
* GO:0006099~tricarboxylic acid cycle 8 genes) (7 w/o outlier)
* GO:0055114~oxidation-reduction process (21 genes) (14 w/o outlier)
* ~~GO:0001649~osteoblast differentiation (8 genes) (not sign. w/o outlier)~~
* ~~GO:0051603~proteolysis involved in cellular protein catabolic process (6 genes) (not sign. w/o outlier)~~
* GO:0002181~cytoplasmic translation (5 genes)
* GO:0015074~DNA integration (6 genes) (5 w/o outlier)
* GO:0045454~cell redox homeostasis (6 genes)
* ~~GO:0006457~protein folding (9 genes) (not sign. w/o outlier)~~
* GO:0006979~response to oxidative stress (8 genes) (7 w/o outlier)
* ~~GO:0051262~protein tetramerization (5 genes) (not sign. w/o outlier)~~
* ~~GO:0006379~mRNA cleavage (4 genes) (not sign. w/o outlier)~~
* ~~GO:0043171~peptide catabolic process (4 genes) (not sign. w/o outlier)~~
* ~~GO:0006397~mRNA processing (11 genes) (not sign. w/o outlier)~~
* ~~GO:0042744~hydrogen peroxide catabolic process (5 genes) (not sign. w/o outlier)~~
* ~~GO:0006508~proteolysis (13 genes) (not sign. w/o outlier)~~
* ~~GO:0015991~ATP hydrolysis coupled proton transport (4 genes) (not sign. w/o outlier)~~
* ~~GO:0046034~ATP metabolic process (4 genes) (not sign. w/o outlier)~~
* ~~GO:0008302~female germline ring canal formation, actin assembly (3 genes) (not sign. w/o outlier)~~
* ~~GO:0032197~transposition, RNA-mediated (3 genes) (not sign. w/o outlier)~~
* GO:0006108~malate metabolic process (3 genes)
* ~~GO:0007293~germarium-derived egg chamber formation (5 genes) (not sign. w/o outlier)~~
* GO:0008340~determination of adult lifespan (10 genes, w/o outlier)
* GO:0070050~neuron cellular homeostasis (3 genes, w/o outlier)

## Jan 16, 2022 & Jan 18, 2022

Generate multiple gene count matrices using various featureCounts settings using the script [featureCounts\_2022-01-18.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/featureCounts_2022-01-18.sh). Summarized results in the [RCK RNASeq Stats](https://docs.google.com/spreadsheets/d/1mk7_pVpPXBE9GQ0E7fPpp1PyzkgYOPV291PJwoviTIM/edit?usp=sharing) spreadsheet tab “featureCounts-summary”

Started exploring other alignment options: wrote slurm script to align RNASeq data to Blue king crab genome using Bowtie2 with RSEM, which is one of Giles’ preferred methods. Waiting for RSEM to be installed on Sedna.

Received blast results from Giles for the Blue king crab genome. Saved blast results to Google Drive in my [P. platypus folder](https://drive.google.com/file/d/14lnN9Kg0b-ZEXO2r1pIxZq6YijewoiPZ/view?usp=sharing), and to the git repo. His notes:

“Ok, so the BLASTing for sequences that Laura gave me finished. It took approximately 6 days to BLAST the 28287 sequences against the nr and nt databases. 5879 (or ~ 20.78%) of the sequences did not find a match to anything within my criteria ( E-Value <= 1E-5 ). I've attached the spreadsheet to this email and here is a breakdown of the columns.

ID Sequence ID from the reference

BlastType BestBlast Hit, basically did the hit originate from the nr (X) or nt (N) database

SeqLength Length of the sequence being blasted

HitAcc Accession number of the hit

HitDesc Description of the hit

Score Score of the alignment between the query and hit

E.Value E-Value of the alignment between the query and hit

PercentIdent The percent identity of the alignment (how well did they match)

QueryStart The base pair point were the alignment starts for the query

QueryEnd The base pair point were the alignment ends for the query

HitStart The base pair point were the alignment starts for the hit

HitEnd The base pair point were the alignment ends for the hit

Note: I used the default bestblast bad word list which consists of Hypothetical, putative, unnamed, unknown and novel protein.”

## Jan 19, 2022

Began re-aligning to BKC genome using RSEM+Bowtie2 using script [align-bowtie2-RSEM\_2022-01-18.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/align-bowtie2-RSEM_2022-01-18.sh)

Backed up raw data (that has been concatenated by tank+crab+read) to Google Drive using:

rclone --drive-shared-with-me --progress --tpslimit 10 --fast-list copy /share/nwfsc/ggoetz/red\_king\_crab/illumina/ remote:RKC\_RNAseq/raw-data/

## Jan 24, 2022

RSEM+Bowtie2 alignment completed.

Started re-aligning using STAR with the script [align-STAR\_2022-01-19.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/align-STAR_2022-01-19.sh)

Some notes from my research on RNASeq bioinf. pipelines:

* RSEM+Bowtie2 is best used when aligning to transcriptome (not applicable here).
* STAR aligner is often used when there is a genome reference, as it is splice-aware

## Jan 25, 2022

* STAR finished. Glanced at results, and alignment rates are a bit lower than Bowtie2 - on average 73.8% +/-2.1%, with 8.1% not assigned due to multi-mapping (mapping to >10 loci), and 15.2% of reads not mapping b/c they were “too short”. The default STAR setting requires that 66% of the read length maps (summed across both mates for paired-end), otherwise it’s tossed. Could ease this requirement, but I’d worry about accurate alignment. I will re-run STAR and adjust the multi-mapping parameters- it has a cut-off of 10, but why not include all multi-mapping loci? Although according to [this thread](https://groups.google.com/g/rna-star/c/mkooNLzyJYc) and [this thread](https://www.biostars.org/p/336962/) STAR does not count any multi mapping reads, which is odd. Biggest issue is that the gene quantification step didn’t work - some issue with the “gene\_id” labels in the GFF3 I think. I need to convert my GFF to a GTF before restarting STAR or something (see [issue 689](https://github.com/alexdobin/STAR/issues/689) and [issue 387](https://github.com/alexdobin/STAR/issues/387)).

## Jan 28, 2022

Used my personal computer to quickly convert the EVM.out\_new.gff3 file to GTF format (EVM.out\_new.gtf) using cufflinks:

/Applications/bioinformatics/cufflinks/gffread EVM.out\_new.gff3 -T -o EVM.out\_new.gtf

Uploaded to Google Drive, then transferred to Senda using rclone:

rclone --progress --tpslimit 10 --fast-list copy remote:EVM.out\_new.gtf .

Updated STAR alignment slurm script to use the GTF formatted annotation file, and I increased the multimapping hit to 50 (but this probably won’t do anything, see above). Started STAR job.

## Jan 29, 2022

STAR finished early afternoon, successfully generated counts! Transferred log files and count files to repo in the /results/star/star.v2/ directory. Most pertinent files are (for example):

* Tank\_9\_Crab\_4.Log.final.out: alignment summary stats
* Tank\_9\_Crab\_4.ReadsPerGene.out.tab: read counts

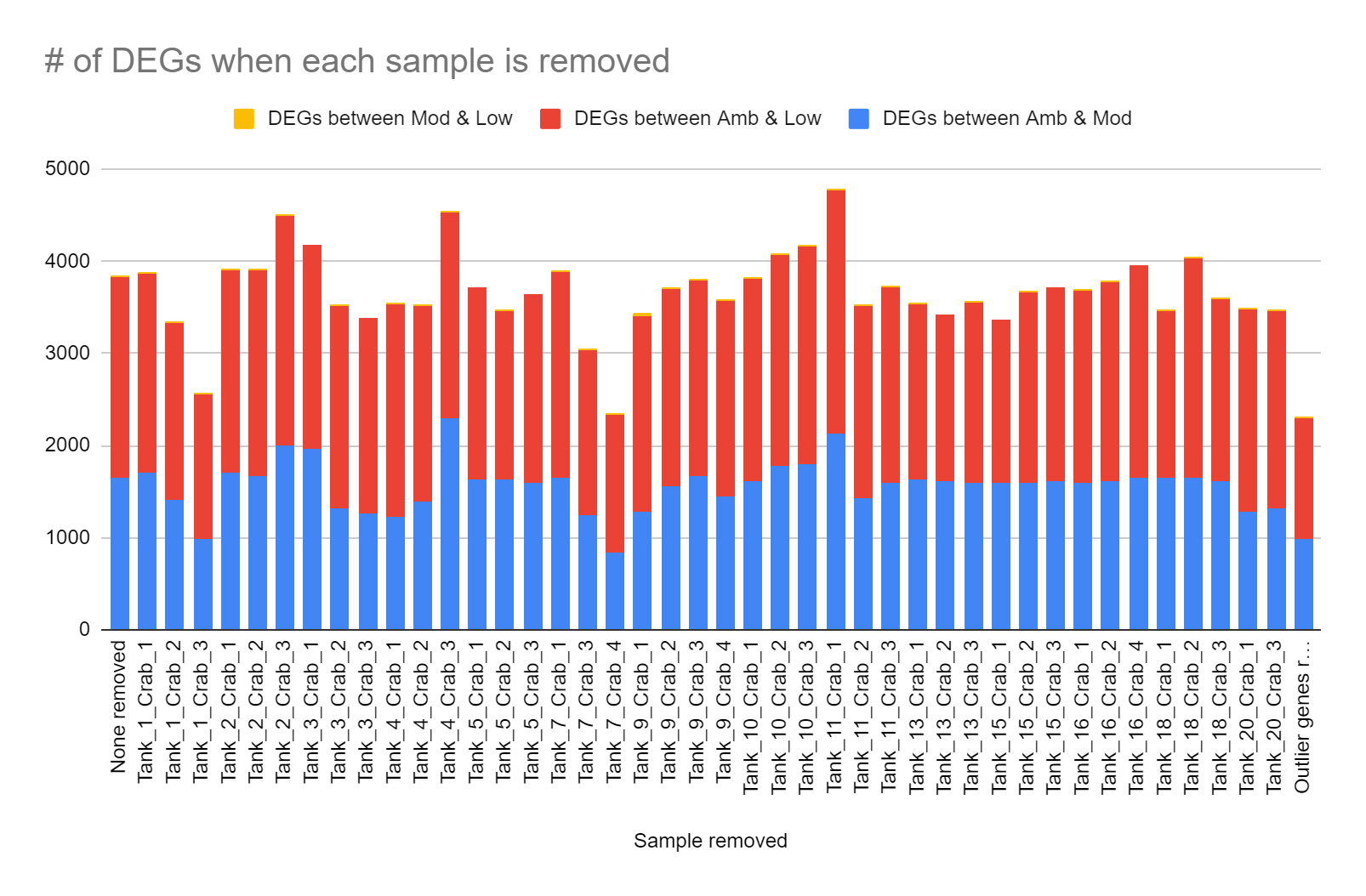
## Feb 8, 2022

Re-ran featureCounts using the Bowtie2 mapped reads to identify the unassigned reads, and inspect chimeric fragments mapping locales (which chromosomes/scaffolds?). It’s a mix - some map to two different chromosomes (1-104), some to a chromosome + scaffold (105+), and oddly others map to the same chromosome. Chimeras represent 1.5% of reads, so to be conservative I will NOT retain chimeras for downstream analysis. I am, however, trying to figure out why some that map to the same chromosome are labeled as chimeras.

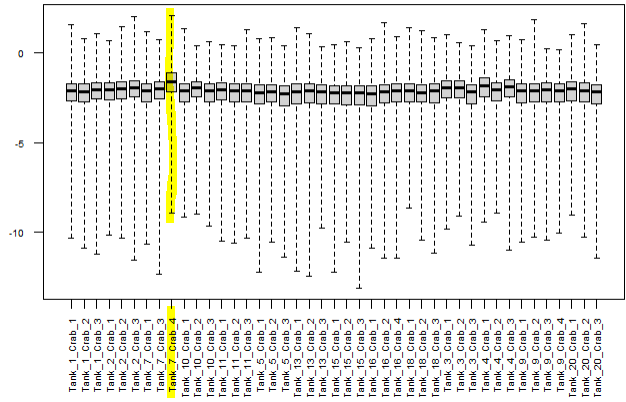
## Feb 9, 2022

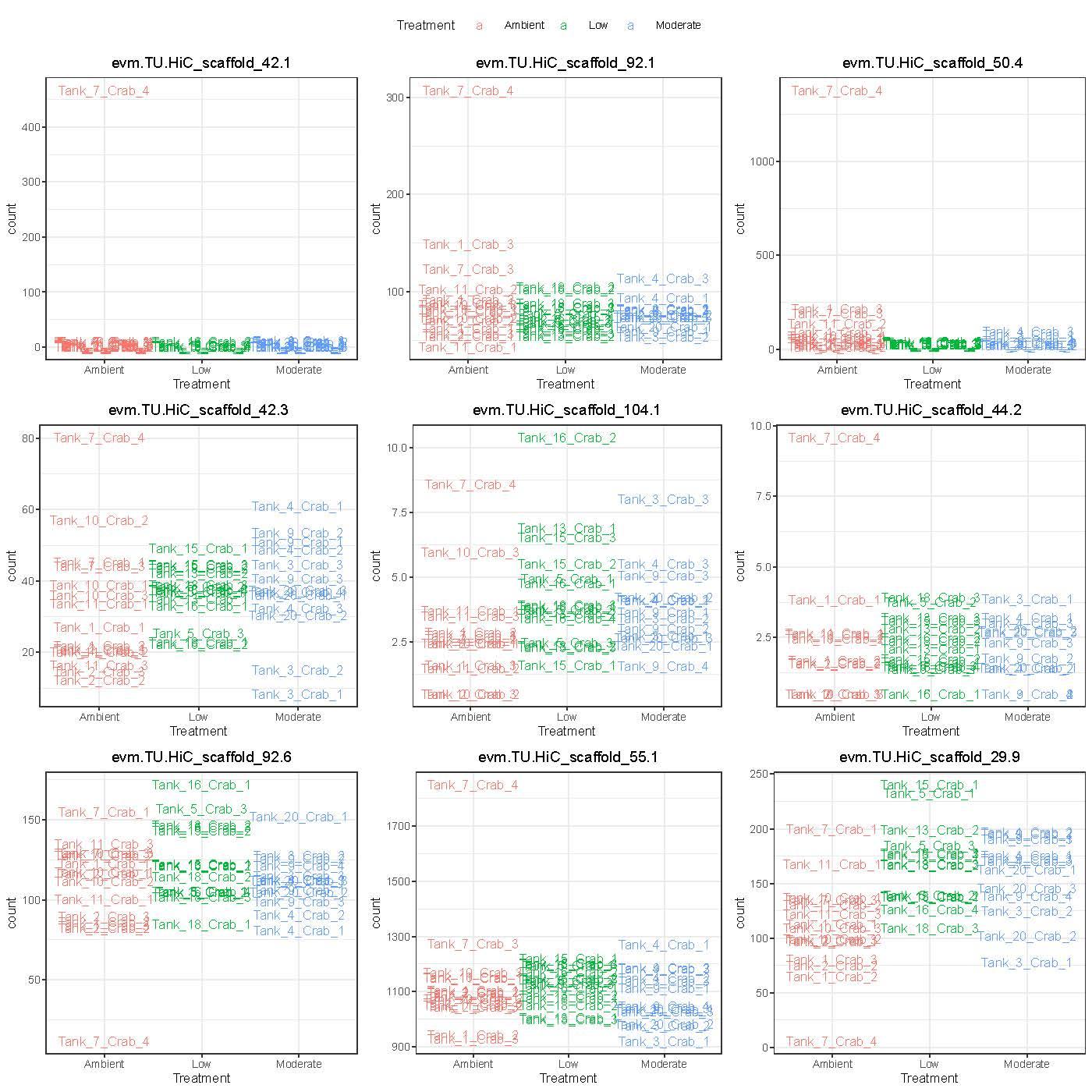
Explored data for outliers. Tank\_7\_Crab\_4 is suspect according to the PCA.

Ran iterations of DESeq using the various featureCounts versions, and did a leave-one-sample-out loop to examine how individual samples impacted DEGs. Also examined the genes for which Tank\_7\_Crab\_4 has outlier counts using DESeq2’s Cook’s Distance statistic. Here is a barplot showing the # of DEGs between each treatment (colors) on the y axis by which sample is removed from the dataset on the x axis. Rightmost bar reflects DEGs when the ~3,000 outlier genes identified in Tank\_7\_Crab\_4 are removed. Tank\_7\_Crab 4 does look slightly more influential, although one could make a case for Tank\_1\_Crab\_3.



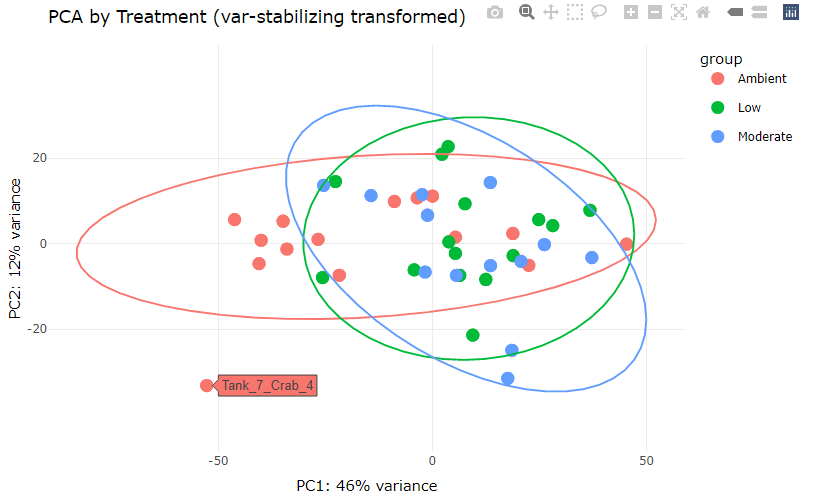
Here are boxplots of Cook’s distances calculated for each gene/sample, with Tank\_7\_Crab\_4 highlighted. Tank\_1\_Crab\_3 doesn’t stand out:





Using the car package’s Boxplot() function I identified the 2,967 genes in which Tank\_7\_Crab\_4 was an outlier. I examined gene count plots for some of those - generally they seemed to indeed be outliers, but there were some that also had similar expression in one or two other samples (often Tank\_7\_Crab\_3). Here’s a random set of 4 genes (of the 2,967), Tank\_7\_Crab\_4 is in red (ambient treatment) and is typically the sample with the highest expression - see also some more on github with better resolution [some-outlier-genes.pdf](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/deseq2/some-outlier-genes.pdf):

I tested removing those outlier genes from the count matrix and re-running the analyses. Tank\_7\_Crab\_4 still was an outlier in the PCA. It really seems like Tank\_7\_Crab\_4 has a unique expression pattern that is not easily controlled for by removing some genes.



**DECISION: Remove Tank\_7\_Crab\_4 - OUTLIER SAMPLE.**

I think that either removing sample Tank\_7\_Crab\_4 OR the outlier genes could be defensible, as they result in similar effects to the number of DEGs. However since Tank\_7\_Crab\_4 still pops up as an outlier on the PCA it seems that there really is something different about that sample.

Interestingly that sample has the largest distinction along PC axis 2 (which explains less variance) - I could inspect the genes that contribute to that axis to see what might be happening? **Related** **Question:** Can we identify the sex of each crab - which chromosome is the sex chromosome(s)? (How even is sex determined in crab?!) - are there transcriptional biomarkers of early sex differentiation in crab that we can use to ID sex?

Ran the STAR aligned gene counts through my data analysis notebooks. Very similar PCAs, numbers of DEGs, etc. I’m summarizing results from the Bowtie2 and STAR alignments in the [RCK RNASeq Stats](https://docs.google.com/spreadsheets/d/1mk7_pVpPXBE9GQ0E7fPpp1PyzkgYOPV291PJwoviTIM/edit?usp=sharing) spreadsheet.

|  | **Bowtie2 (featureCounts v3)** | **STAR** |
| --- | --- | --- |
| **Mean alignment rate** | 78.44% | 73.77% |
| **No. of genes (across all samples)** | 28,287 | 28,286 |
| **Ave. no. of genes detected in each sample** | 20,398 | 16,518 |
| **Total gene counts, all samples** | 606,506,395 | 502,075,954 |
| **No. genes filtered due to low frequency (<10 across all samples)** | 5,939 | 9,786 |
| **No. genes remaining for DESeq2 analysis (across all samples)** | 22,348 | 18,500 |
| **DEGs, Amb vs. Moderate** | 833 | 630 |
| **DEGs, Amb vs. Low** | 1,469 | 933 |
| **DEGs, Moderate vs. Low** | 19 | 19 |
| **Enriched Biological Processes - Amb vs. Moderate** | 8 | 8 |
| **Enriched Biological Processes - Amb vs. Low** | 21 | 15 |
| **Enriched Biological Processes - Moderate vs. Low** | 1 | 1 |

## Feb 10, 2022

Added a third column to the above table to reflect the Bowtie2-mapped data but NOT including singletons, as I wondered whether that gene set would better match the STAR-mapped data (i.e. in the # of DEGs, enriched processes). Not really. I’m troubled by the fact that Bowtie2 and STAR produce different results (varying numbers of DEGs, enriched BP). Should decide which method to use *without* considering the results (i.e. do not cherry-pick). Alternatively- I could come up with a consensus list of DEGs - but do people do that?

Started incorporating Giles’ Better Blast results into my annotation and enrichment analyses. Problem is that the blast results don’t include GO terms or gene IDs that can easily be used in enrichment analyses. Tried pulling the NCBI accession numbers as-is, and without the version (i.e. removed the “.#” from accessions) and using in DAVID but they weren't recognized. Also tried translating NCBI accession numbers to UniprotIDs and GO terms using various online tools (e.g. <https://www.uniprot.org/uploadlists/>), but didn’t work. It seems that there is a mix of accession number formats/types, such as:

* MPC15113.1
* XP\_045116590.1
* KAG7162023.1
* LR584238.1

How to proceed?

**To do**: should filter the Best Blast results for lower e-value. But what should be the e-value threshold?

## Feb 15, 2022

Explored consensus DEGs identified using both Bowtie2 and STAR. Overlaps aren’t great.

|  | Bowtie2 | STAR | Consensus |
| --- | --- | --- | --- |
| No. genes in DESeq2 analysis | 22,348 | 18,500 | 18,326 |
| DEGs, Amb vs. Moderate | 833 | 630 | 376 |
| DEGs, Amb vs. Low | 1,469 | 933 | 633 |
| DEGs, Moderate vs. Low | 19 | 19 | 9 |

The consensus set would be pretty small. I worry that STAR throws out all multi-aligned genes during the gene quantification step (while Bowtie2 finds the best alignment), and I’m not sure how STAR handles singletons. For gene exploration and discovery, which is the purpose of this study, I will go with Bowtie2 as my aligner.

##### SUMMARY OF DATA PROCESSING DECISIONS:

* Use Bowtie2 aligner
* Generate counts using featurecounts: count fragments (not reads), do not include chimeras, do include singletons, use default overlap length (1 bp)
* Remove outlier sample (Tank\_7\_Crab\_4)
* Do not remove any genes identified as outliers for particular samples.
* Do not remove fragments mapping to genes on scaffolds (105+).

##### Further analysis notes & ideas:

Filtering low-frequency genes:

* Should I do this? If so, do I remove those with mean OR sum count less than ten across all samples.
* Or should I first normalize gene counts THEN identify and remove low-frequency genes? [(Koch et al. 2018)](https://paperpile.com/c/X1tIZE/N4hT) says yes. What threshold to use?
* DESeq2 automatically filters low-frequency genes, replacing the p-values in the results object with NA. Therefore, no real need to do this! So why do other folks suggest it? Some genes that are borderline DEGs could pop up if I reduce the starting gene set.
  + Decision 3/15: I remove low-frequency genes before DESeq2 analysis (other papers do this, e.g. <https://doi.org/10.1186/s12864-017-4392-0>). I retain genes that have mean count of 10 across all samples OR genes with counts>=10 across at minimum 10% of the samples.
  + I ran DESeq2 without removing any genes, and cross-checked the DEGs with all genes removed due to low counts, and there are no genes that are tossed out at the filtering step that could then be DEGs. So I feel good about this step.

Plotting counts: should first normalize gene counts before visualizing. Does DESeq2 do that?

* Yes. “normalizes counts by the estimated size factors”.

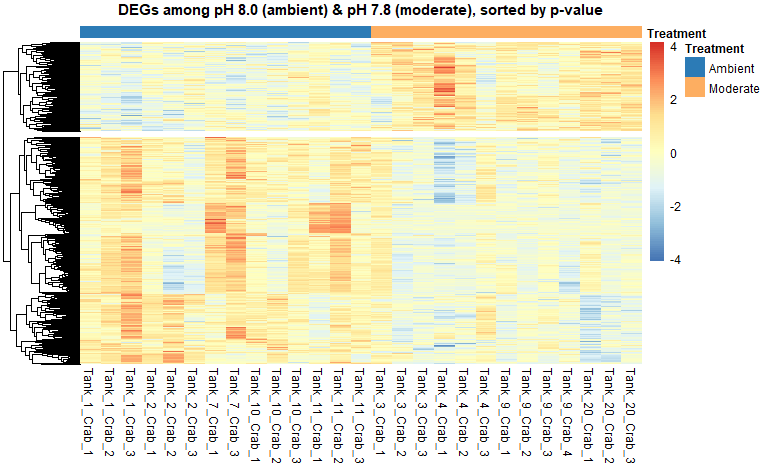
Tank effects?

* Broadly from PCA- doesn’t appear to be a big issue. DESeq2 can’t include tank as a covariate in analyses- but other software can. Perhaps try [MDSeq](https://github.com/zjdaye/MDSeq).
* When I cluster samples using gene counts (e.g. in heat map) or pairwise sample distances, they do not cluster by tank. That is good.
* Various PC axes - any separate by tank?

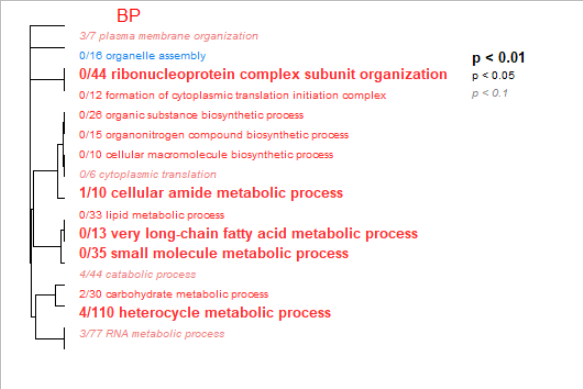
How to explore variance heterogeneity?

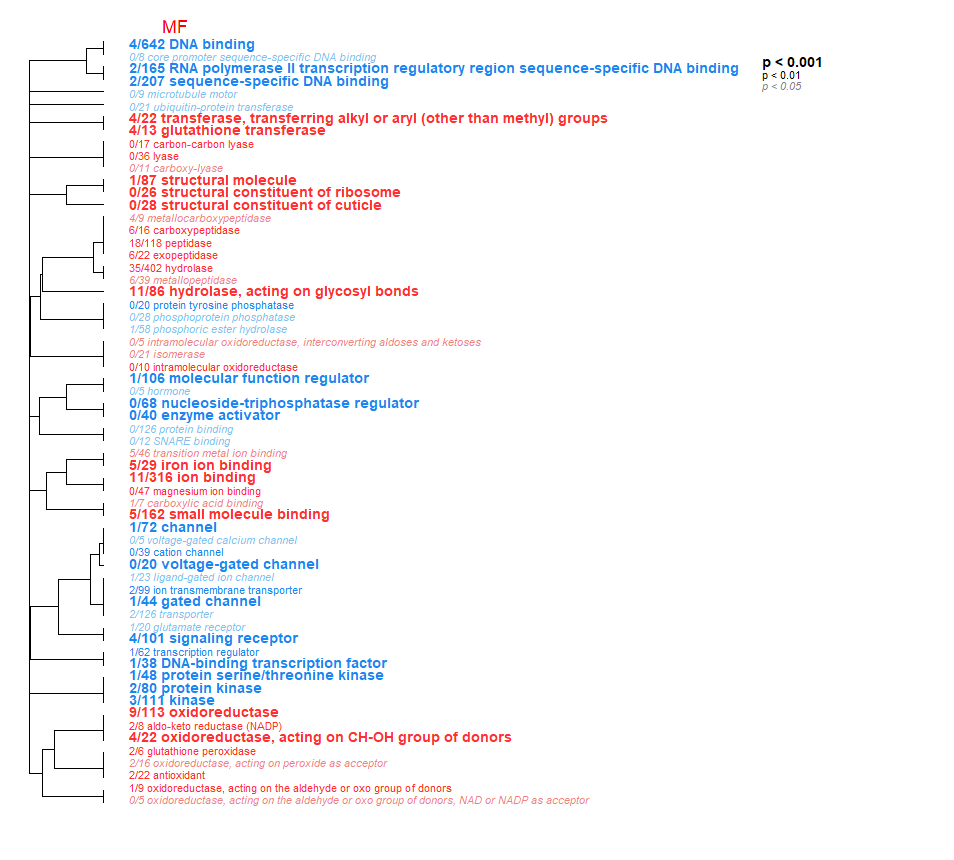
* Broad differences - distance to centroid in PCAs using a) all genes and b) DEGs.
  + Hypothesis: variability in expression of stress-response genes under stressed conditions could be beneficial to the population as a whole. Bet hedging. Under normal conditions, likely less variability in stress-response genes, and more variability in other processes.
  + Idea: plot PCA of DEGs vs. PCA of all other genes - does the within-treatment spread change?
* Gene-based variance:
  + Coef. Variation for each gene- look at distributions by treatment.
    - When to calculate CV? Probably after normalization (but which norm.? DESeq2’s internal?)

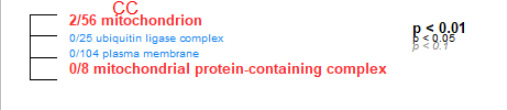
#### Differential Expression Analysis summary

There are 820 differentially expressed genes in ambient pH vs. moderate pH. Of those, 589 are less abundant in moderate pH, and 31 are more abundant. Here’s a [list of annotated DEGs](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/DEGs.Amb-vs-Mod_annotated.csv). 

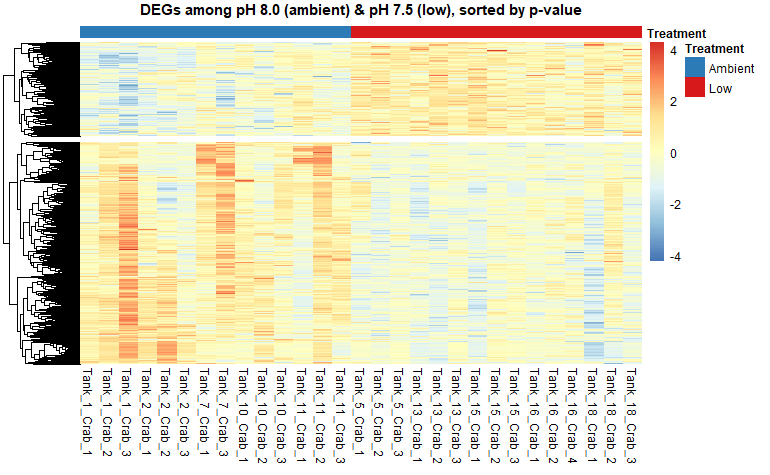
Enrichment analysis using GO\_MWU: 16 biological processes, 72 molecular functions, 4 cellular components: red=enriched in ambient, blue=enriched in moderate pH



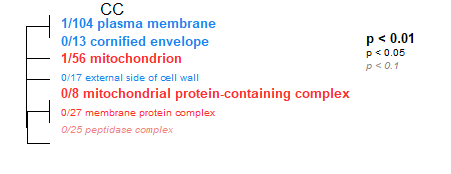
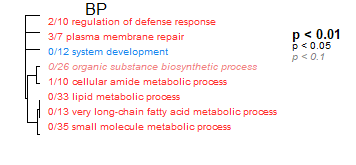


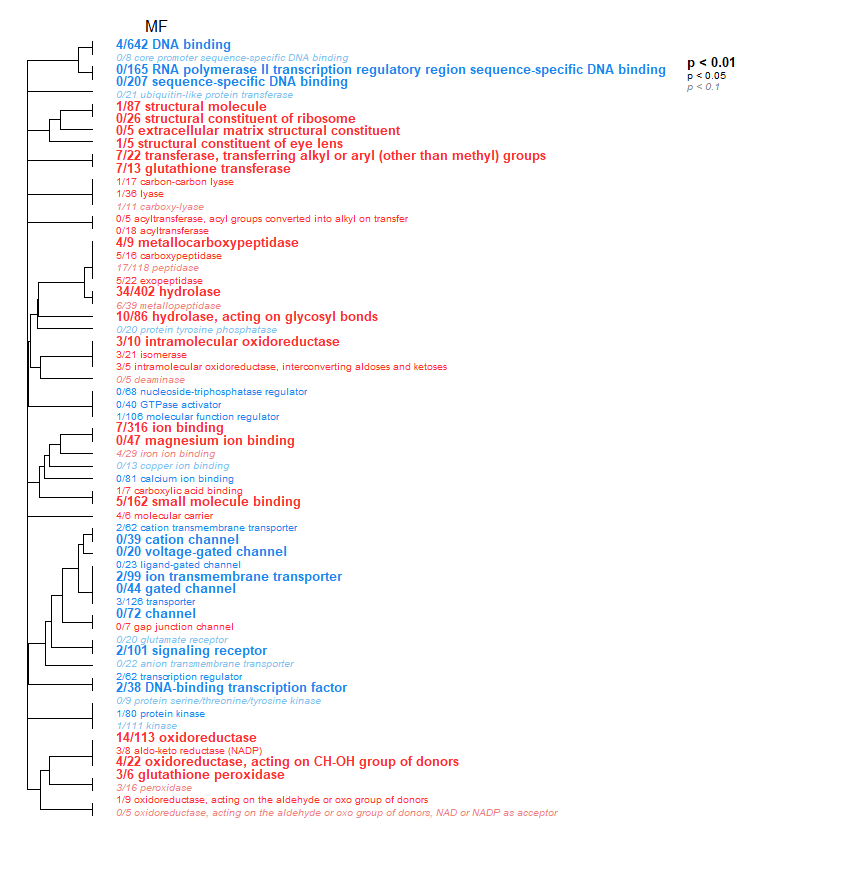


There are 1,464 differentially expressed genes in ambient pH vs. low pH. Of those, 1,030 are less abundant in low pH, and 434 are more abundant. Here is a [list of the annotated DEGs](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/DEGs.Amb-vs-Low_annotated.csv).

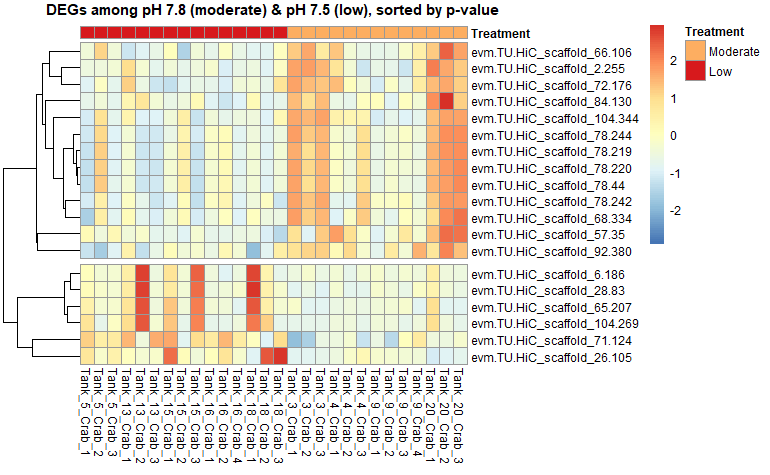


Enrichment analysis using GO\_MWU: 8 biological processes, 63 molecular functions, 7 cellular components: red=enriched in ambient, blue=enriched in low pH

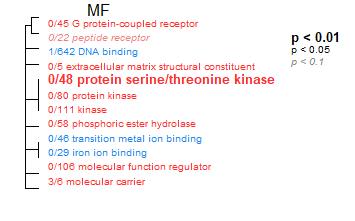


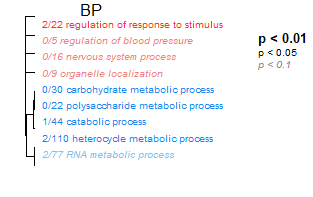


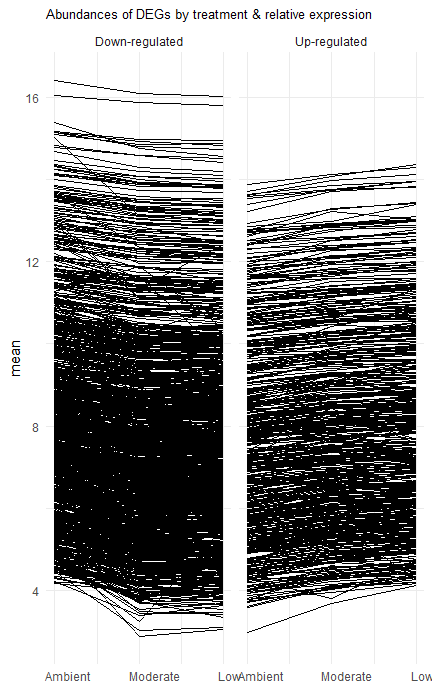
There are 19 differentially expressed genes in ambient pH vs. moderate pH. Of those, 13 are more abundant in moderate pH, and 6 are more abundant in low pH . Here is a [list of these DEGs](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/DEGs.Mod-vs-Low_annotated.csv).



Enrichment analysis using GO\_MWU: 9 biological processes, 12 molecular functions, 0 cellular components: red=enriched in moderate pH, blue=enriched in low pH

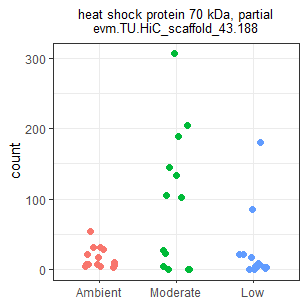
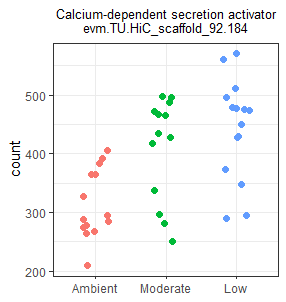
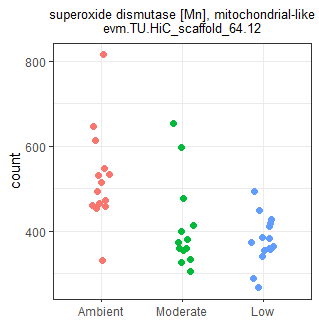






Question: Do genes that respond to OA change “linearly” with pH (i.e. in a dose-dependent manner)? The plot on the right shows per-treatment mean of normalized gene counts (y-axis) for all DEGs by ambient, moderate, and low pH treatment on the x-axis. They are separated by the DEGs that are less-abundant (“down-regulated”, left plot) and more-abundant (“up-regulated”, right plot) as compared to the ambient pH treatment. It certainly does appear that yes, expression decreases with OA severity in the downregulated genes, and similarly increases with severity in the upregulated genes. (Investigate this more using WGCNA analysis)

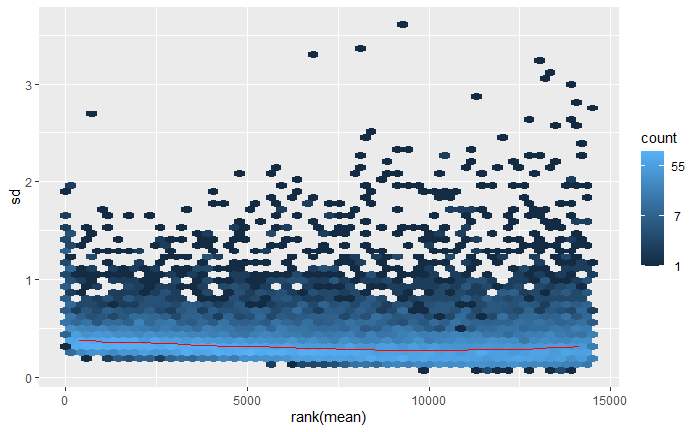
I became interested in this question, because I was inspecting some DEGs, and observed a ~linear change in some gene counts by pH. For example, check out these two genes, superoxide dismutase and calcium-dependent secretion activator.



But not all genes fit that trend. For example see this heat shock protein gene:

## Feb 16, 2022

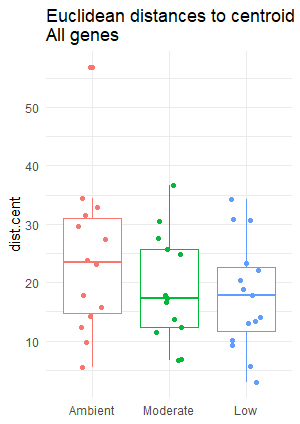
##### Variance analyses

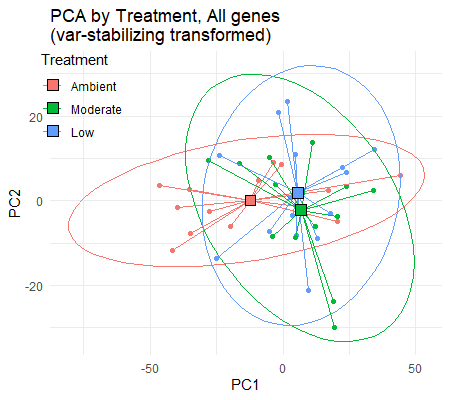
Explored global differences in dispersion among treatments using PCAs. 

First, PCAs were constructed from gene counts that had been variance-stabilized (using DESeq2). This smooths out differences in variance associated with the number of counts (i.e. typically high variance at lower counts), but is intended to preserve other sources of variation (i.e. experimental differences). To the right is a plot showing per-gene variation (sd) by its ranked mean value after the variance-stabilizing transformation.

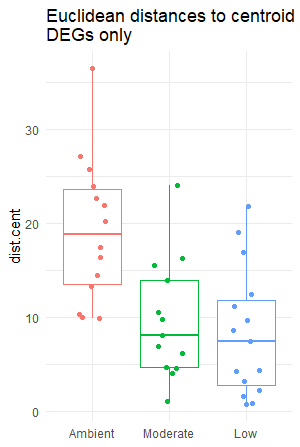
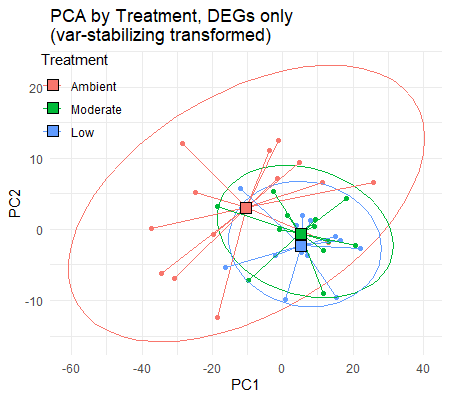
Using the PC1 and PC2 coordinates for each sample, I calculated centroids for each treatment (mean of all PC1 and PC2 coordinates), then the euclidean distances of each point to its treatment centroid (sqrt(a2 + b2) = c). I compared per-sample distances to the centroid using boxplots and ANOVA (after testing/transforming for normality). I then replotted the PCA biplot with the addition of each treatment’s centroid.

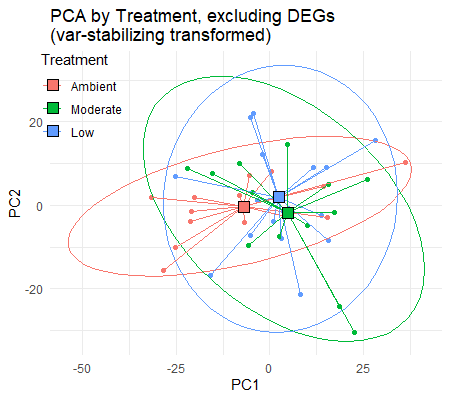
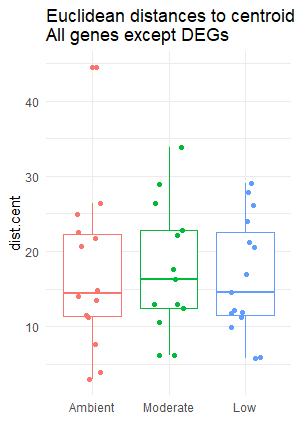
Here is the PCA with centroids, and boxplots of euclidean distances (from each sample to the centroids) for ALL genes analyzed. Distances do not differ significantly among treatments (F(2,39)=1.00, P=0.38):





Here is the PCA with centroids, and boxplots of euclidean distances using ONLY the genes that were differentially expressed (DEGs). Distances differ among treatments (F(2,39)=10.1, P=0.00029), and Tukey post-hoc tests confirm that the euclidean distances are larger in the ambient treatment vs. moderate (p-adj=0.0052) and vs. low (p-adj=0.0034):

Finally, here are the same figures using all genes EXCEPT the genes that were differentially expressed, representing genes that are not responsive to treatment. Distances did not differ among treatments (F(2,39)=0.054, P=0.95). 

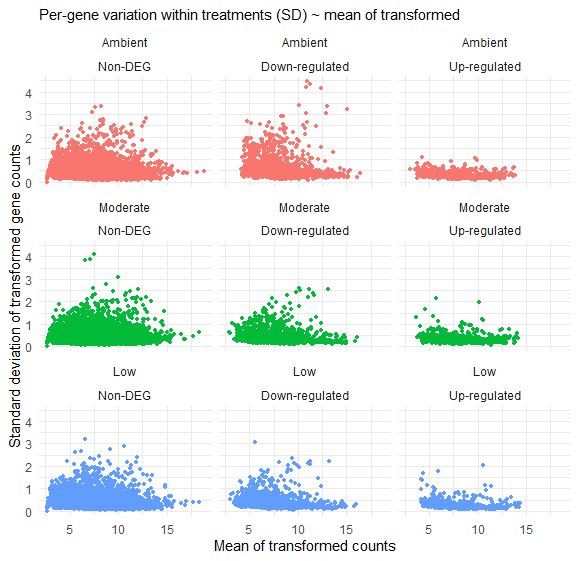


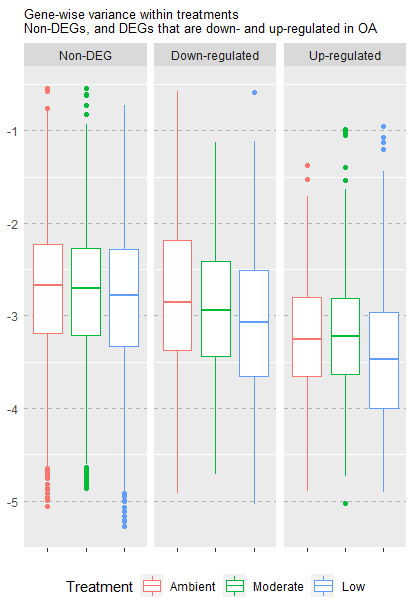
My takeawayS:

We observed from the all-gene PCAs that the crab held at ambient seem to have a larger overall variability in gene expression. The question is- is this variability across the board (all genes), or is it specific to genes that are responsive to OA (aka differentially expressed)? From this analysis, the answer seems to be that specific genes that respond to OA (either upregulated or downregulated) are also less variable across individual crabs. So, it appears that some of the differential expression *could be* driven by large within-group differences in the ambient crab - they are able to be themselves (unique!), rather than conform to a critical set of physiological processes that are required by the acidified environment.

New questions that are arising:

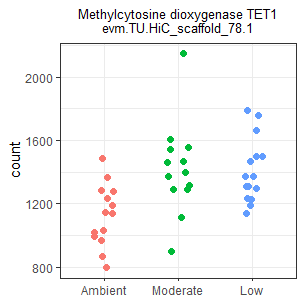
* Does the pattern observed above hold when we look at per-gene variances? Need to compare per-gene variance by treatment for All genes, DEGs, and non-DEGs.
* Are genes that are up-gregulated in OA more or less variable? I need to identify genes that are more abundant in OA. Same question with genes that are downregulated.
* Since PCAs distill many variables into two axes, it’s possible that a small set of genes is influencing the PCA - which genes are top contributors to PC1 and PC2? How many contributing genes is required for the differences in variance to emerge?

I calculated the mean, standard deviation, and coefficient of variation for each gene (using normalized counts) within each treatment. In the **left figure** below, I’m showing the coefficients of variation of transformed counts for non-DEGs, and DEGs that are down-regulated and up-regulated in either OA-treatment compared to ambient. I excluded the handful of genes that were only diff. expressed between moderate & low pH treatments. Note that I log-transformed the CV for visualizing. In the **right figure** I show the standard deviation plotted against the mean for each gene within each treatment, for genes that are down-regulated, up-regulated, and unresponsive to OA treatments. 



Main take-aways:

* Genes down-regulated in OA conditions are less variable among individuals compared to crabs in the ambient treatment. This group of genes are therefore not only less abundant, but also less variable.
* Genes that are up-regulated in OA conditions have similar variability among crabs held in moderate vs. ambient. There does seem to be less variability in gene expression in the low pH crabs (blue), possibly indicating that those upregulated genes are critical to function & survival at very low pH, and regardless of the individual they need to be abundant at particular levels.
* In general, differential expression and variability in expression is dominated by OA inducing less activity (fewer genes are upregulated in OA conditions, this is also obvious in the heatmaps). Would be interesting to have metabolic rate information. Lit review?

Which genes were consistently expressed at higher levels in OA conditions? Here is a list of genes that were upregulated in OA (either low or moderate pH) with very little inter-individual variability (CV<0.03): [DEGs.up-low-variability\_annotated.csv](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/DEGs.up-low-variability_annotated.csv). **These genes possibly represent those that are critical for functioning in OA-acclimated crab.** NOTE: these genes have been annotated twice, first with the Best Blast results, second with my Uniprot/Swissprot blast results. In some cases, they differ. Best to look at the gene with the lowest e-value. As I look over the list, many appear to be transcriptional regulation & transposable element activity. Interestingly, Methylcytosine dioxygenase TET1 (right) is upregulated in both the low and moderate pH treatments at consistent levels, indicating higher demethylation activity in OA treatments. 

There are a few upregulated genes (more abundant in OA treatments) that are also highly variable in moderate & low OA treatments, as indicated by the outliers in the boxplots. I wonder what those are? To answer that question, I filtered for DEGs with sd>1 that are upregulated in OA treatments, and inspected their Best Blast and my Uniprot/Swissprot blast annotations, and list those below alongside their mean+sd of transformed counts (one gene is omitted as it is not annotated):

| Gene ID | Description | NCBI accession | Moderate pH mean + SD | Low pH mean + SD |
| --- | --- | --- | --- | --- |
| evm.TU.HiC\_scaffold\_57.349 | cuticle protein CP1876-like [Penaeus japonicus] | XP\_042874395.1 | 5.42  1.34 | 5.92  1.77 |
| evm.TU.HiC\_scaffold\_43.188 | heat shock protein 70 kDa, partial [Munida rugosa] | ANQ44686.1 | 5.69  2.13 | 4.26  1.37 |
| evm.TU.HiC\_scaffold\_6957.5 | mannose-binding protein [Penaeus vannamei] | ROT65260.1 | 4.61  1.66 | 4.345  1.67 |
| evm.TU.HiC\_scaffold\_21.268 | methionine aminopeptidase 1-like isoform X2 [Homarus americanus] | XP\_042228484.1 | 10.43  1.20 | 10.78  1.17 |
| evm.TU.HiC\_scaffold\_68.177 | PREDICTED: Penaeus monodon histidine-rich glycoprotein-like | XM\_037948726.1 | NA | 5.37  1.04 |
| evm.TU.HiC\_scaffold\_2.255 | probable ribonuclease ZC3H12C [Homarus americanus] | XP\_042231575.1 | 8.42  1.06 | NA |
| evm.TU.HiC\_scaffold\_28.83 | Solute carrier family 22 member 2 [Nymphon striatum] | KAG1682452.1 | NA | 4.76  1.13 |
| evm.TU.HiC\_scaffold\_26.219 | Thymelicus sylvestris genome assembly, chromosome: 9 | OU426894.1 | 3.69  1.30 | 4.12  1.41 |
| evm.TU.HiC\_scaffold\_26.102 | Trichoderma reesei strain CBS999.97 chromosome VI, complete sequence | CP020878.1 | 10.60  1.15 | 11.35  1.12 |
| evm.TU.HiC\_scaffold\_59.460 | uncharacterized protein LOC113825471 [Penaeus vannamei] | XP\_027234108.1 | 10.18  1.99 | 10.52  2.02 |

*Possible additional analyses to assess heterogeneity of variance:*

* I need to find a statistical test to compare gene-wise variances among treatments that is robust to large sample sizes - when I use anova(CV ~ Treatment\*DEG) all factors and pairwise comparisons significantly influence CV, similarly I see differences when I test for homogeneity of variance using Levene’s and Fligner tests on the normalized count data.
* Perform linear regression of normalized gene counts ~ mean gene counts, pull the residuals, then perform ANOVA of those residuals by treatment, up/down. This was done in a paper - find which one.

## Feb 17, 2022

##### Genetic composition analysis

Question: what is the genetic composition of test crab? Did selection occur during treatments that resulted in varying genetic composition among treatments? Could this be influencing the greater variation in gene expression that we see in ambient treatment?

Wrote a slurm script (based on old script from O. lurida project) to call variants from sorted alignment files (.bams), [genotype-from-RNASeq.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/genotype-from-RNASeq.sh). Started a test job on just two files to see if the pipeline works.

## Feb 26, 2022

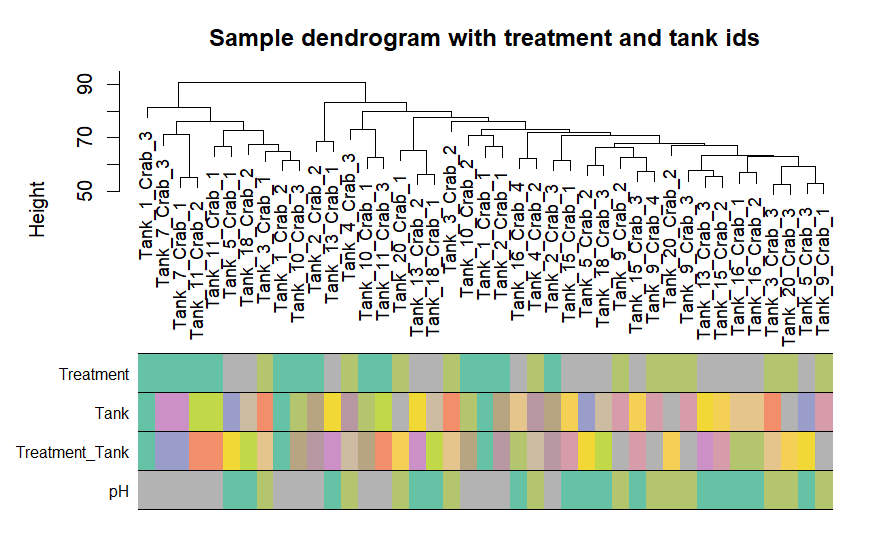
Genotyping job terminated due to time-out (took way longer than I thought!). It stopped during the variant calling step (#6). Took the easy route to keep things moving along by moving the .bam files that were already used to produce .g.vcf’s to new directory, then began the job starting at the `HaplotypeCaller` step. Did not move the .bam for sample Tank\_13\_Crab\_2 b/c the job terminated while processing it (probably not complete).

Tried to use the new multi-threaded version of HaplotypeCaller to speed things up (`gatk HaplotypeCallerSpark`), but it didn’t run since it’s still in beta-mode.

NOTE: My genotyping script adds the same read group to all samples for gatk’s purposes. They were each, however, split across multiple lanes. Since reads were concatenated by sample into one .bam, I couldn’t specify the lane at this point. I would have to go back and align the reads separately for each non-concatenated fastq file. I think the current script makes sense, since all samples were sequenced on the same run, all split across all lanes, so any per-lane variation should apply to all samples. Plus, I’m using genotype information for a course assessment of relatedness, not to identify variants of interest.

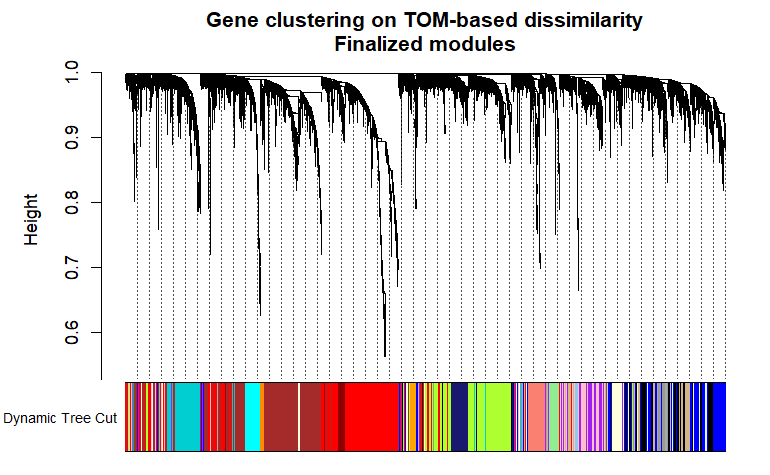
## Mar 3, 2022

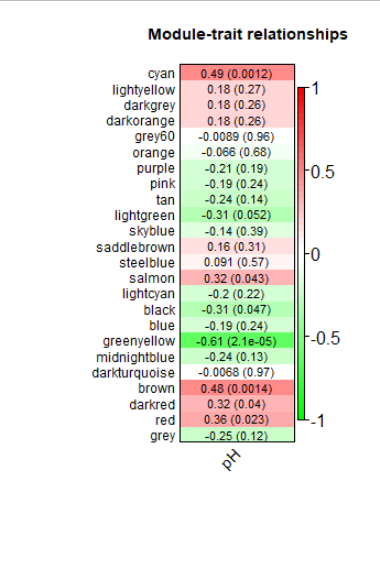
#### WGCNA Analysis

Ran data through a [Weighted gene co-expression network analysis](https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/#abstract) (WGCNA). This is an unsupervised method of finding clusters (modules) of highly correlated genes and relating them to traits/treatments (e.g. pH). In essence, you feed the software all your genes (after removing low-frequency genes and outlier samples), and it calculates a series of distances/correlations among genes, groups them, then correlates their pattern to the trait of interest. The full analysis is in an RMarkdown notebook, [03-WGCNA.Rmd](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/notebooks/03-WGCNA.Rmd), but here are some key figures. 

One lingering question I had was if there were pronounced tank effects. Here is a dendogram, the results of hierarchical clustering by sample, which shows that samples largely don’t cluster by tank (which is good!).

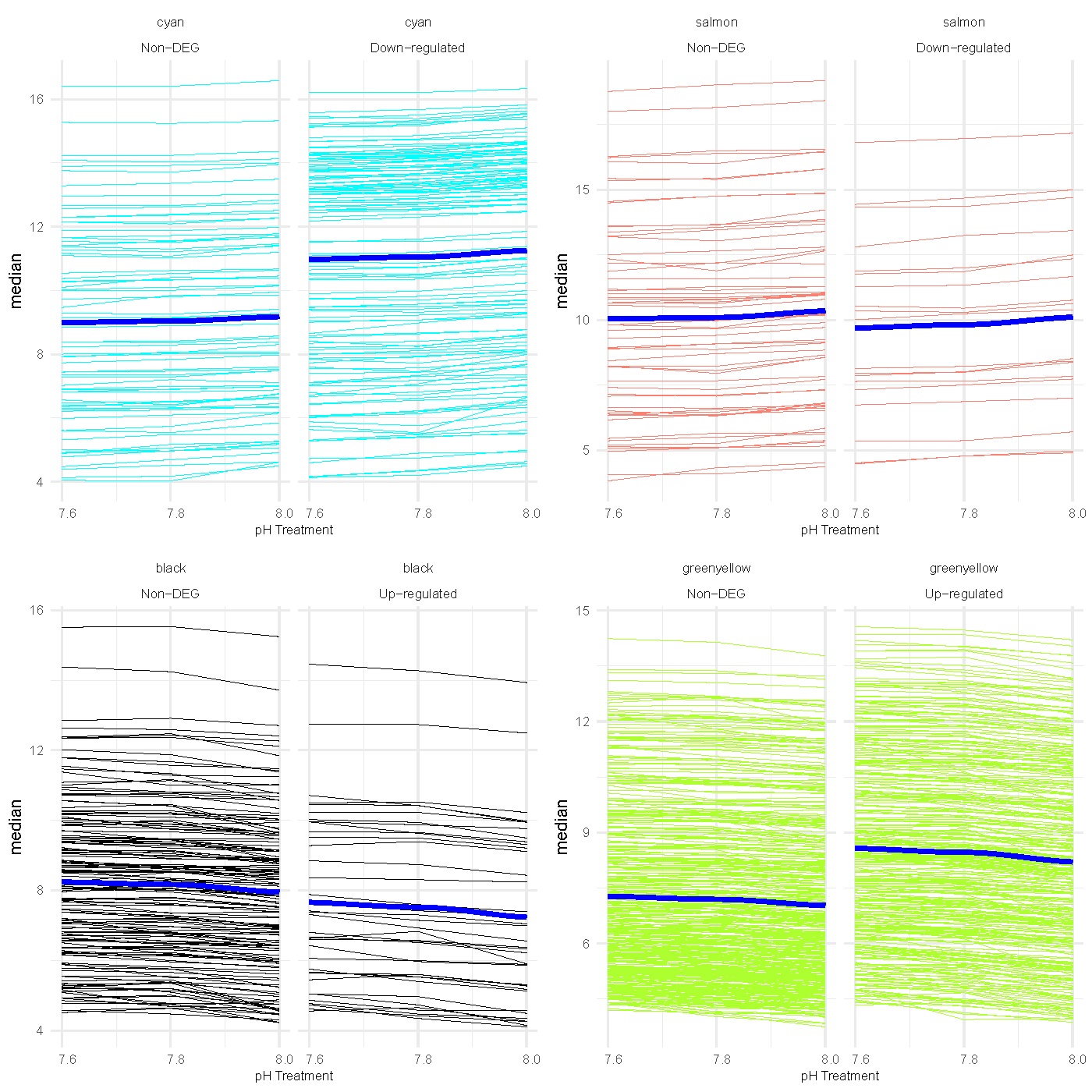
Now for a summary of WGCNA module results. WGCNA first calculates per-gene adjacencies (a reflection of co-expression similarity), then transforms adjacency into Topological Overlap Matrix, calculates dissimilarity, then clusters based on dissimilarity. The result is this dendrogram showing how genes cluster. The color bands on the bottom indicate the modules to which each gene is assigned. “Branches of the dendrogram group together densely interconnected, highly co-expressed genes.”

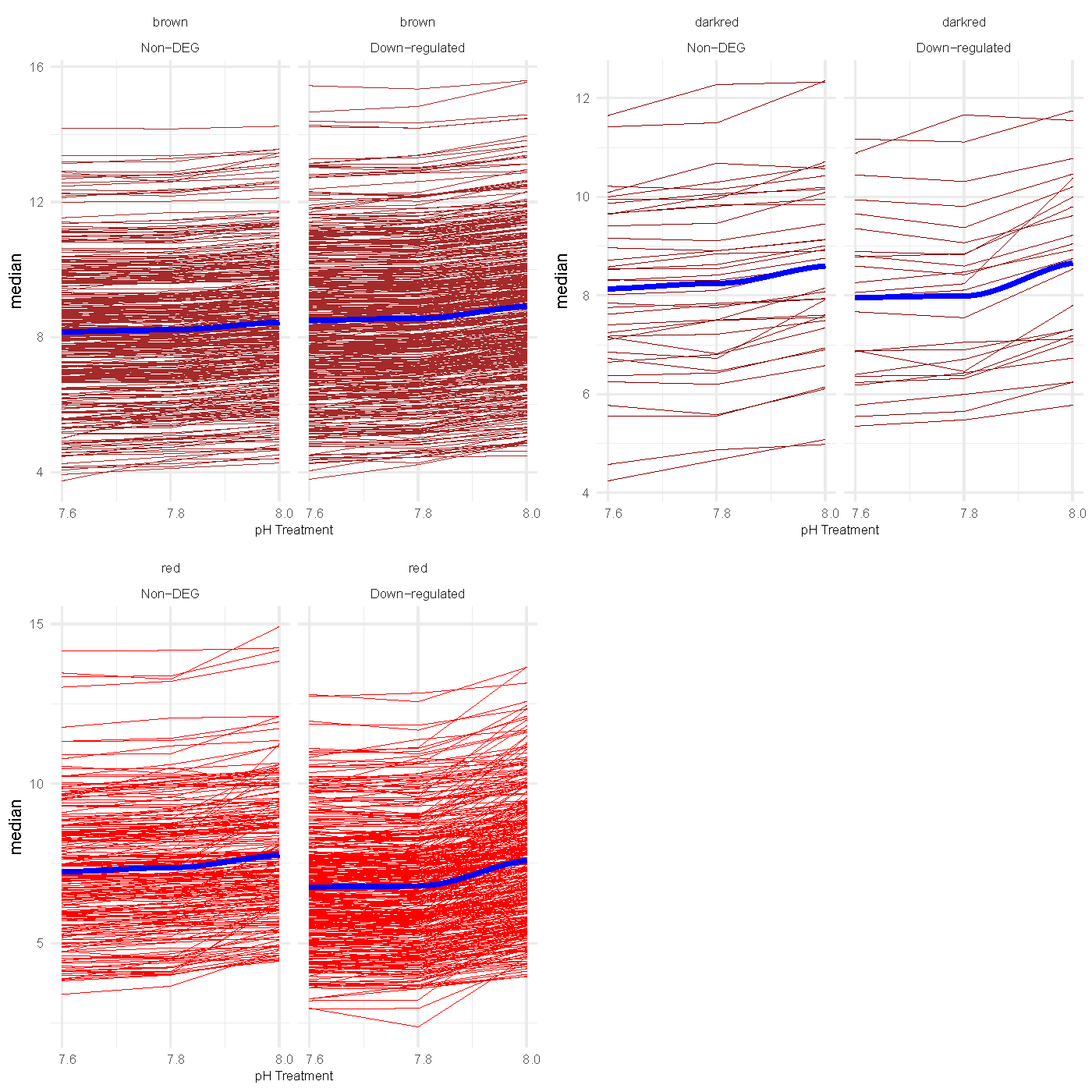




After the gene modules are identified, they are examined for correlation with a trait of interest, which in my case is pH. From the WGCNA tutorial, “*We quantify associations of individual genes with our trait of interest (weight) by defining Gene Significance GS as (the absolute value of) the correlation between the gene and the trait.”* The resulting Module-trait relationship figure (right) indicates the correlation between gene modules and pH. Each row corresponds to a module eigengene. Each cell contains the corresponding correlation and p-value. Given alpha=0.05, genes in the cyan, salmon, brown, darkred, and red modules are positively associated with pH, while genes within the black and greenyellow modules are negatively associated with pH.

But how do genes respond with pH in each module? I generated line plots for each module, using only the genes that are significantly associated with pH (gene significance p-value (p.GS.pH) <0.05). Below are those plots - each line represents one gene, and the line connects three points - the median expression (normalized) for each pH treatment. For each module I separated genes into two groups, genes that are & are not differentially expressed between ambient and OA treatments according to DESeq2 (Non-DEG, Downregulated, or Upregulated). Remember, colors represent different modules (co-expressed genes). I added blue trendlines (loess) to each plot to show the overall response of each modules’ genes to pH. For higher quality figures, see [pH-associated-modules-genes.pdf](https://github.com/laurahspencer/red-king_RNASeq-2022/blob/main/results/wgcna/pH-associated-modules-genes.pdf).



****

My takeaways-

* There are several groups of co-expressed genes that are associated with pH.
* Most modules (cyan, salmon, brown, darkred, red) indicate that gene abundances decrease as pH conditions are reduced. Two modules contain genes that increase in activity as pH decreases.
* It’s very encouraging that there is consistency between WGCNA and DESeq2. Lots of DEGs (from DESeq2) are contained within the pH-associated modules (from WGCNA), and no modules contain both upregulated and downregulated genes.
* The WGCNA analysis is probably a better way to approach the enrichment / network analyses - i.e. I need to examine the overall biological functions of each module. I can use DESeq2 results to inform interpretation, such as saying that module “greenyellow” contains genes that are more active in OA (indicated by the high number of downregulated DEGs in that module).
* Question to answer: What % of DEGs are captured in pH-associated WGCNA modules? Answer: 78.5%.
* Perhaps a heatmap would better illustrate patterns in these modules?

## Mar 7, 2022

##### Struggling with NCBI accession numbers for enrichment analyses

Returned to the NCBI accession ~ GO term issue for enrichment analyses. I successfully matched NCBI accession numbers to the NCBI GI numbers using the [NCBI Batch Entrez tool](https://www.ncbi.nlm.nih.gov/sites/batchentrez). This tool requires uploading .txt files of accession numbers for protein and nucleotide accession numbers, separately. I therefore did the following (using the excel file “best-blast-GIs.xlsx”, for some reason- if I have to do it again I’ll use R):

* Split best blast results by BlastType: blastx and blastn
* Filtered each of the blastx and blastn tables to exclude those with e-value > 1e-10. I did this b/c the batch entrez tool seemed not to handle the larger sets of genes.
* Removed the version # from each accession #.
* Saved gene sets to .txt files (one gene per row)
* Uploaded files one at a time to the Batch Entrez tool:
  + Best-blast\_blastx\_evalue-max-1e10\_accessions.txt, selected the database “Protein”
  + Best-blast\_blastn\_evalue-max-1e10\_accessions.txt, selected the database “Nucleotide”
* After the tool finished, I clicked on the link to view the results and downloaded a .txt file of GIs, then also downloaded a .txt file of accession numbers, via: Send To -> File -> GI or Accession. (annoyingly no download option included both, so I simply copied/pasted them into a spreadsheet, and manually confirmed that they were in the same order).

I uploaded the NCBI GI numbers to the Uniprot ID conversion tool - but unfortunately only a small fraction mapped to Uniprot IDs:

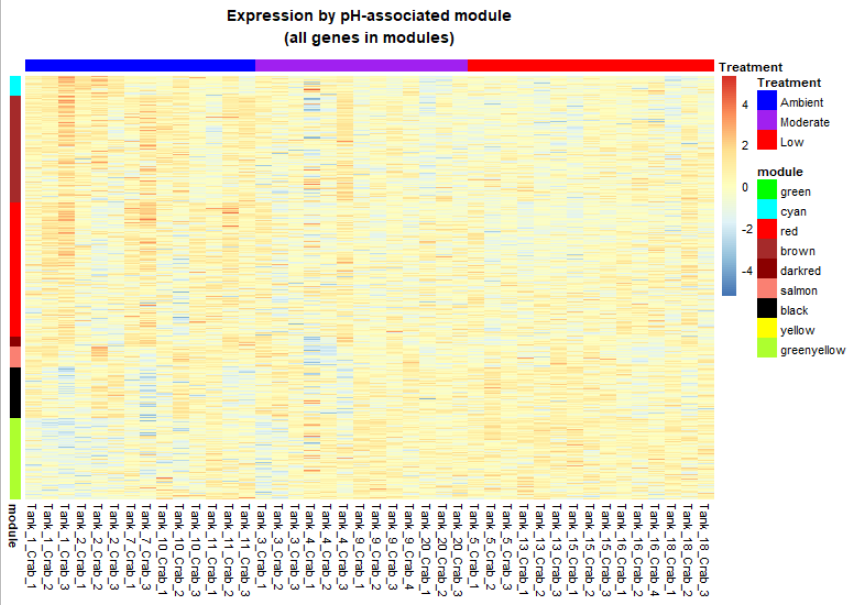
* 664 out of 14605 GI number identifiers were successfully mapped to 665 UniProtKB IDs in the table below. Out of the remaining ids, 270 were successfully mapped to the UniParc sequence archive.

I also tried using the NCBI accession numbers as my input (using blastx and blastn results separately), with very low hits:

* 243 out of 13883 RefSeq Protein identifiers were successfully mapped to 244 UniProtKB IDs. (Out of the remaining ids, 10495 were successfully mapped to the UniParc sequence archive)
* 11 out of 722 RefSeq Nucleotide identifiers were successfully mapped to 11 UniProtKB ID.

**Very frustrating!!!** Perhaps the best course of action is to use my Uniprot Blast table - or re-blast using a slightly less stringent e-value threshold for a higher annotation rate. I posted my issue to the [Roberts Lab Github discussion board](https://github.com/RobertsLab/resources/discussions/1419). Sam said that the gene ID conversion tools (including David) don’t work well with a mix of ID types (which I have, since the Best Blast protocol uses multiple databases). **He and Steven both suggested using my own Uniprot blast results** for the purposes of gene enrichment analysis**.**

Mar 9, 2022

Continued working with the WGCNA software - generated heatmaps to show expression patterns for each of the pH-associated modules of genes. They are below. Each row represents a gene, column a sample, and cell coloration represents the transformed counts (variance-stabilization transformation). The genes are organized by module, with the top-most module (green) the most positively associated with pH, and the bottom-most module the most negatively correlated with pH. There are two heatmaps - the top contains all genes assigned to the significant modules, and the bottom contains only genes that are significant “members’ of the modules. 

##### 

## Mar 10, 2022

#### Functional annotation & enrichment analyses

To obtain the name, function and GO terms for each gene, I re-blasted the blue king crab gene sequences (EVM.out\_new.cds) against the Uniprot/Swissprot database with an e-value cut-off of 1e-5, then used [Uniprot’s batch retrieval tool](https://www.uniprot.org/uploadlists/) to get GO terms for each annotated gene. When actually using the annotation table, for genes with multiple hits I selected the lowest e-value. Of the 28,287 total gene sequences described in the blue king crab genome, 14,340 were annotated at e-value 1e-5, and 11,930 were annotated at e-value<1e-10. Thus far, I have used the more liberal e-value cut-off of 1e-5 for my enrichment analyses.

For GO enrichment analyses I also tested the [eggnog-mapper](http://eggnog-mapper.embl.de/) instead to obtain GO terms directly from the cds fasta file, which is [suggested by the GO\_MWU developer](https://github.com/RobertsLab/resources/discussions/1342), but the number of annotated genes was much lower - 4,292.

##### Gene enrichment analysis of WGCNA modules using GO MWU

There are 7 modules of genes (containing from 204 - 2,821 genes, each) that are significantly associated with pH. Using [GO\_MWU](https://github.com/z0on/GO_MWU) I identified enriched biological functions for each module. Following the GO\_MWU instructions for WGCNA results, I used the module membership scores for genes assigned to modules as the test statistic (as opposed to p-value or log2fc). Here are the results for each module:

###### Module = Black

* **Increased** abundances in OA
* 1,086 genes
* No enriched biological processes or molecular functions
* 6 enriched cellular components

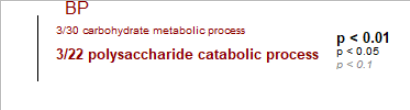
###### 

###### Module = Greenyellow

* **Increased** abundances in OA
* 1,710 genes
* Enriched: 2 molecular function & 1 cellular component (cornified envelope, p<0.01)

*Module = Darkred*

* **Decreased** abundances in OA
* 204 genes
* Enriched: 2 biological processes, 7 molecular functions , & 1 cellular component (extracellular region, p<0.01)



*Module = Salmon*

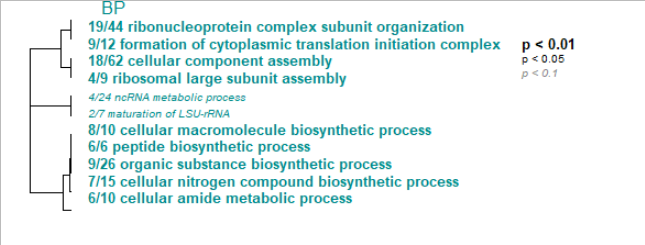
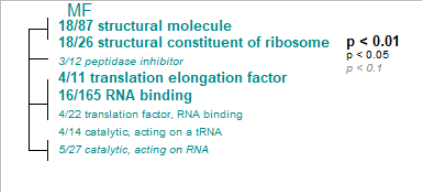
* **Decreased** abundances in OA
* 441 genes
* Enriched: 5 molecular functions, & 2 cellular components

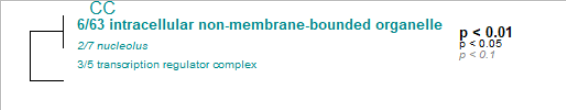


###### 

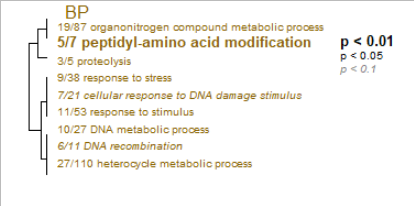
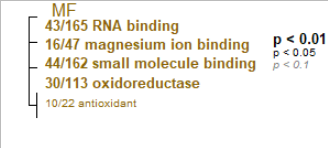
###### Module = Cyan

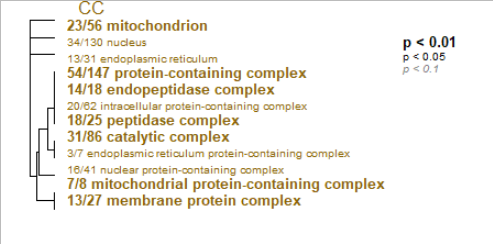
* **Decreased** abundances in OA
* 433 genes
* Enriched: 11 biological processes, 8 molecular functions, & 3 cellular components





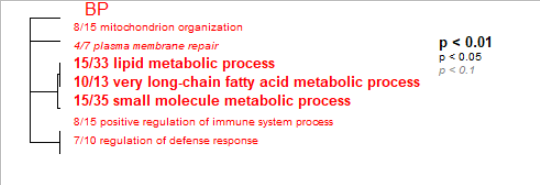
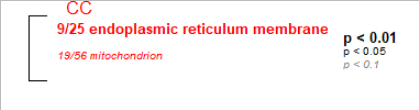
###### Module = Brown

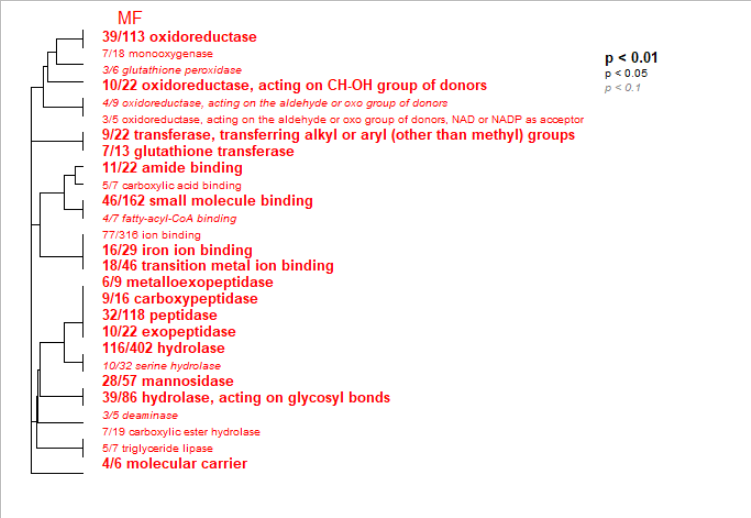
* **Decreased** abundances in OA
* 2,244 genes
* Enriched: 9 biological processes, 5 molecular functions, & 12 cellular components



###### Module = Red

* **Decreased** abundances in OA
* 2,821 genes
* Enriched: 7 biological processes, 27 molecular functions, & 2 cellular components





## Mar 11, 2022

Made some slides with figures of results to date- [RCK summary of results -](https://docs.google.com/presentation/d/1EeemSu5K0gr0NW8qzNM7X9REU-8SnMgZBCHqlf8W9hg/edit?usp=sharing) Mar 11, 2022.

Showed Krista & Ingrid, here are some notes and next steps:

* IPA - send DEG lists and WGCNA results to Giles
  + X Prep DESeq for IPA
  + Prep WGCNA for IPA
* X Generate Venn diagram of DEGs among pH treatment contrasts
* X Generate Venn diagram of DEGs vs. WGCNA pH-associated module membership
* Variance analysis - see Shellys dungeness crab paper for another method of variation comparison ??? didn’t see a variance comparison …
* Look into lit. related to crab (and other species) physiological response to stress.
  + Hypothesis: Under stress conditions - more limiting physiologically, canalizing/forced into certain set of processes.
* Show functional analyses to Chris - he might know a lot about the enrichment analysis stuff
* Selection impact? Clustered genotypes vs. clustered expression.

## Mar 23, 2022

Quick update:

**More WGCNA adjustments:**

Explored the option of starting WGCNA from unfiltered counts (vst-normalized) to identify outlier samples, rather than using the DESeq2 input. This was to examine whether DESeq2 & WGCNA had similar outlier samples. Tank 7 Crab 4 still was a clear outlier. Another sample (Tank 4 Crab 1) also clusters outside the rest of the samples, *however* that sample is less defensible as an outlier given the branch distance on the sample dendrogram (and the PCA & leave-one-out analysis). SO - I decided to only remove the clear outlier (Tank 7 Crab 4), which allows me to use the same count matrix for both DESeq2 and WGCNA. This makes comparing the two analyses simple. If reviewers want to see more evidence for my outlier sample removal, I can always include the sample tree called “Sample clustering to detect outliers” using WGCNA in the supplemental materials.

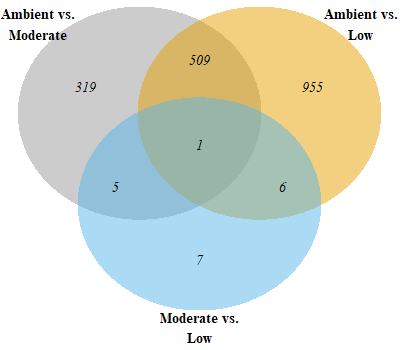
There are two WGCNA settings that are user-specified, and which can potentially impact results, described below. After exploring these settings I re-ran that analysis.

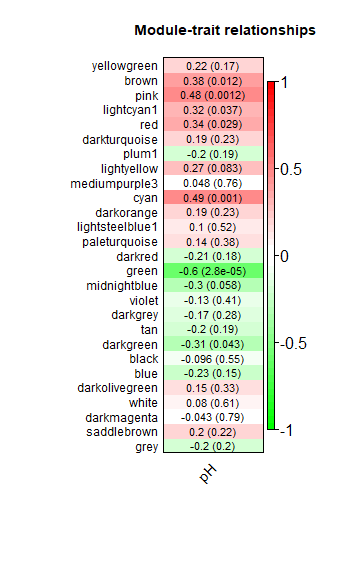
1. Minimum number of genes in each module. The setting used in the WGCNA tutorial is 30 genes. I tested running with 30 and 50 - thinking that a larger module would reduce the parsing of genes. Ultimately the resulting pH-associated modules did not differ, so I decided to set the **minimum module size to recommended 30 genes**.
2. Minimum correlation among eigengeens to use when merging modules. As per the tutorial, “The Dynamic Tree Cut may identify modules whose expression profiles are very similar. It may be prudent to merge such modules since their genes are highly co-expressed. To quantify co-expression similarity of entire modules, we calculate their eigengenes and cluster them on their correlation.”. The tutorial uses a correlation threshold of 0.75. Thinking that I only wanted genes that very similar expression patterns to be merged in to the same module, I explored using higher thresholds, such that highly correlated modules would be merged. I tested using 0.8, 0.85, and 0.9. This did, indeed, change the number of modules associated with pH. However, increasing the correlation threshold also resulted in modules with fewer genes (which makes sense). As I glanced through the lineplots for each module and the heatmap, having more modules doesn’t appear to parse the data into clearer trends - having fewer larger modules seems to be the best way to proceed. Also, when unsure about a setting, I think it’s best to go with the software developer’s suggestion. So, ultimately, **I will use the recommended cut height of 0.25 on the module eigengene correlation dendrogram to merge modules, corresponding to a minimum correlation of 0.75 .**

I developed an RMarkdown notebook for the gene enrichment analysis, and wrote an R function (!) to run through each DEG set and module gene set.

#### Exploring gene set overlaps via Venn Diagrams

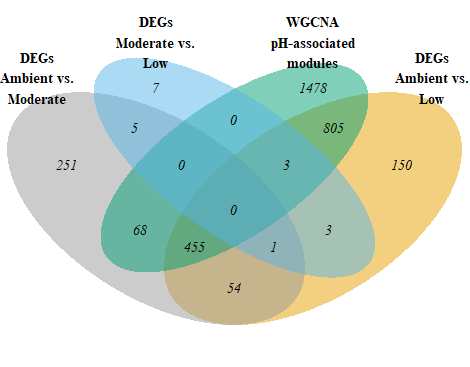
There are 1,802 genes that are differentially expressed in one pH treatment compared to another. Given that 18,150 genes were examined in DESeq2 (~10k were pre-filtered due to low frequency), that corresponds to 9.9% of examined genes that are responsive to pH treatments. Of the 1,802 DEGs, a large portion were differentially expressed between ambient pH and low pH (955, 53%). 319 genes differed between ambient & moderate pH (18%). 509 (28%) were differentially expressed in both OA treatments (moderate and low pH) as compared to ambient.



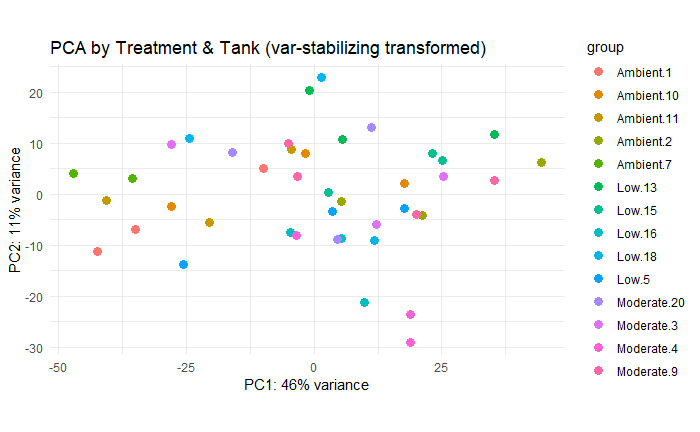
There are 7 modules of genes that are significantly associated with pH treatment: brown, pink, lightcyan1, red, cyan, green, and darkgreen (see “Module-trait relationships” figure, with sign. modules indicated by p-value<0.05, in parentheses). The total number of genes that are assigned to any of these pH-associated modules is 9,298 (that’s 51% of all genes examined in WGCNA). When I filter for only genes that are also individually associated with pH, that drops to 2,809 genes (15.5% of all genes) (i.e., p-value for the Gene Significance (GS), which is the absolute value of the correlation between the gene and pH, is <0.05). 

There are quite a few genes that are identified by both DESeq2 as differentially expressed, and by WGCNA as associated with pH. Of the 1,802 DEGs, 1,466 (81%) were assigned to a pH-associated module (1,331 genes, 73.3%, after I filter for only genes that are also individually determined as significantly associated with pH (p.GS.pH<0.05). That’s quite a bit of overlap!

Here is a venn-diagram showing the overlap among DEGs for each pH contrast, and WGCNA-identified genes that are associated with pH (found in pH-associated modules, and significantly associated with pH on their own).



#### Examined several PC axes for potential tank effects



Note:

* Could look into trial run for IPA - 2 week free

Enrichment analysis to use

* Steve Horvath TEs
* Ask BKC authors for TE feature track

TO DO:

* Multivariate homogeneity of variance test - betadisper from vegan package. Refer back to multivariate stats course materials
* TE analysis:
  + What type of TEs were retained in BKC genome?
  + Reach out to BKC authors for TE annotation file
  + Plan of attack for TE analysis
* Genotype analysis- PCA and correlation with expression PCA
* IPA - review tutorials / manual
* Write methods & sketch out results
* Lit review
  + Crab stress physiology
  + Crab juvenile physiology

#### Mar 25, 2022

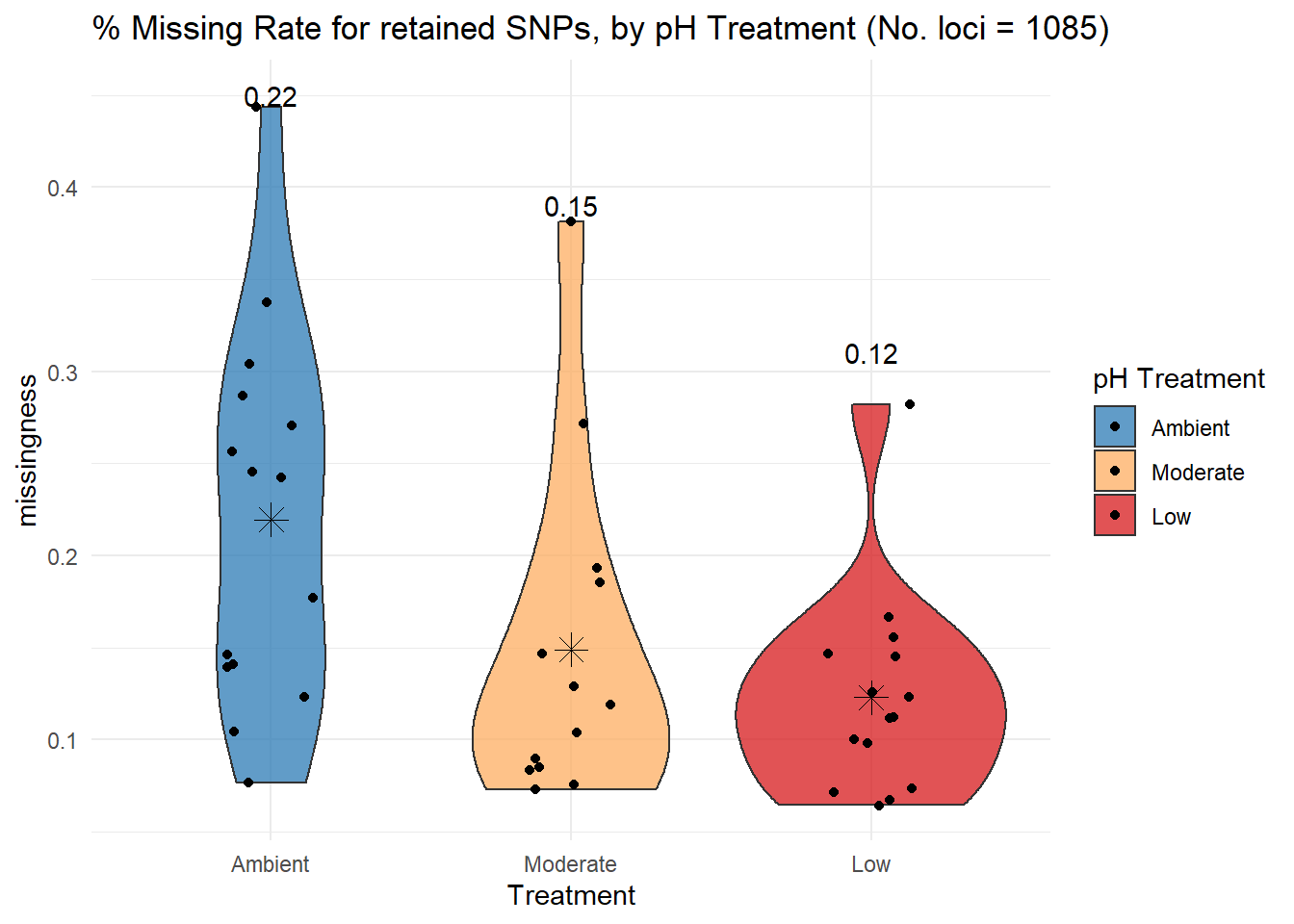
##### Running Ingenuity Pathway Analysis

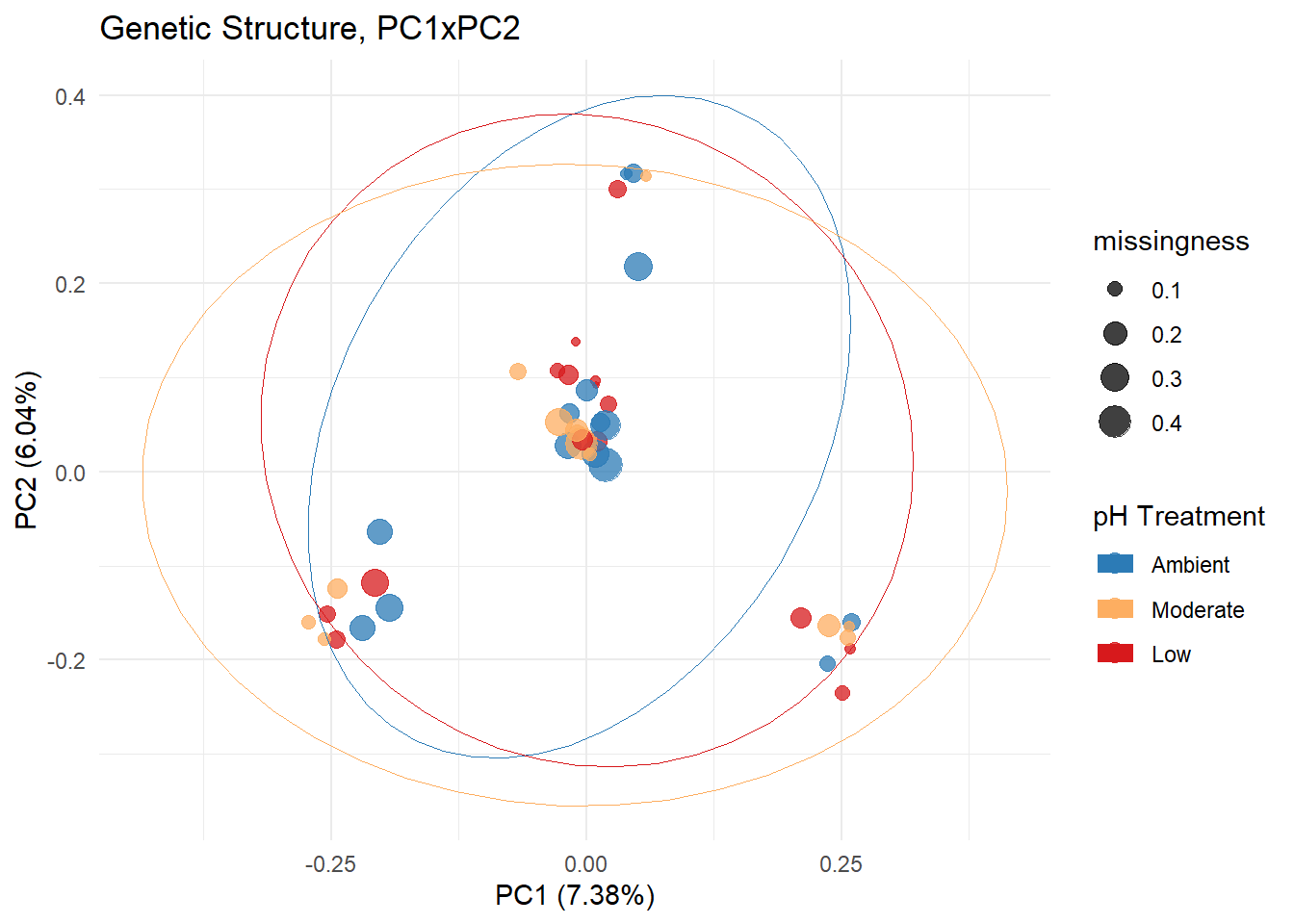
Met with Krista & Giles to run gene sets of interest through IPA. Gene sets include 1) DEGs for each pH treatment contrast, 2) Gene Significance

## Apr 6, 2022

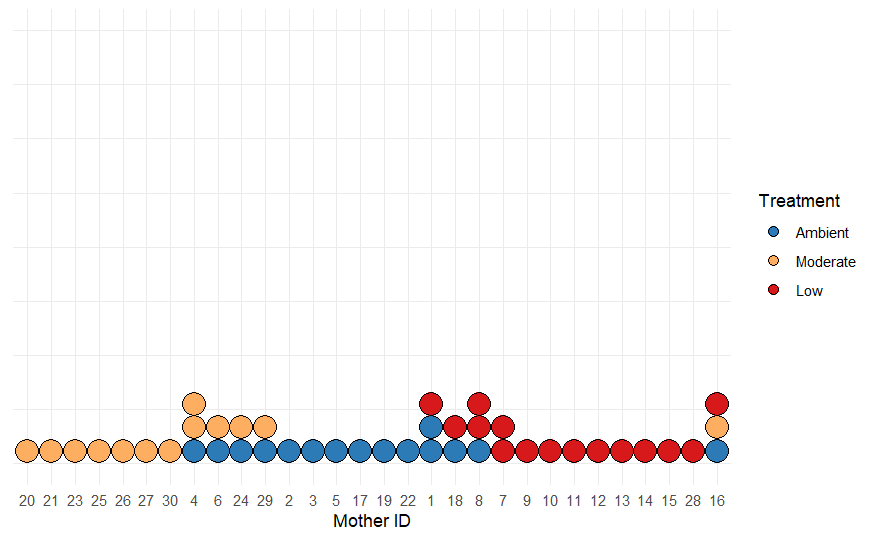
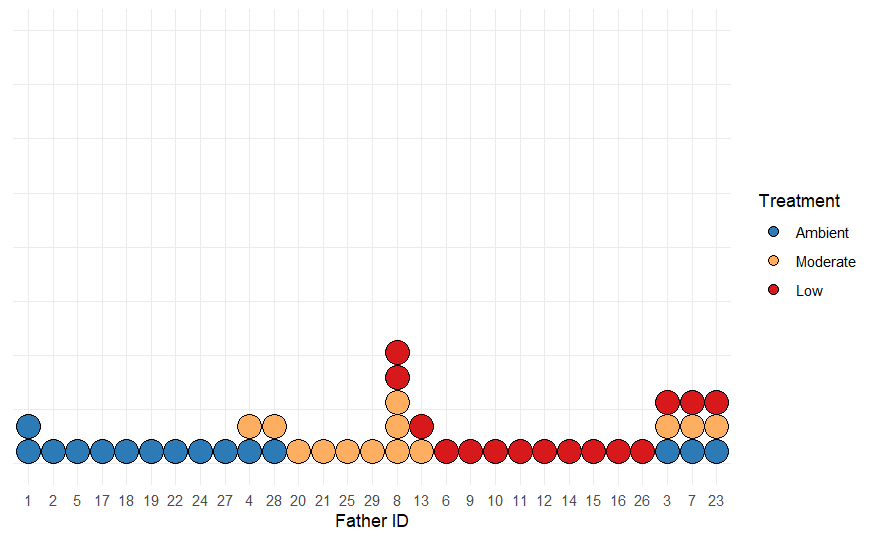
#### Summary of genetic analysis

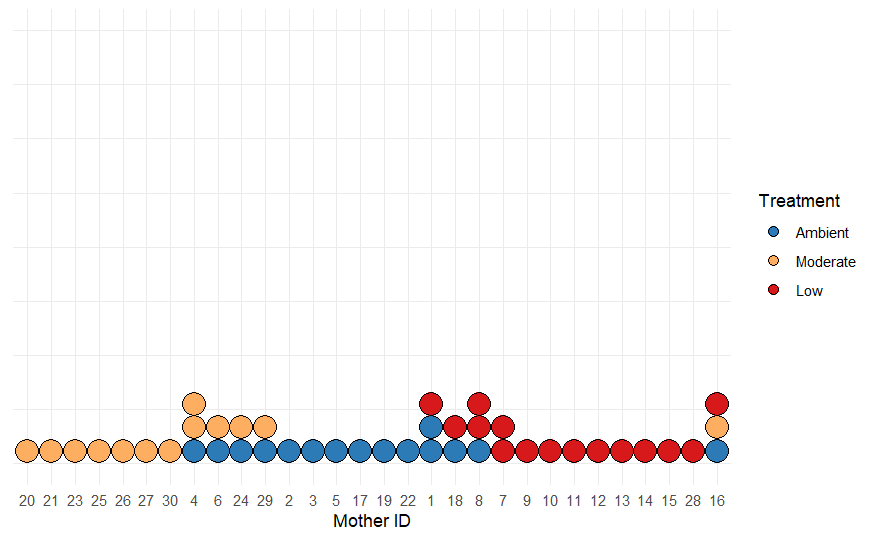
To Do - Note, decided to change max % missingness to 25%. This didn’t change clusters, but did result in more outlier loci and a few that are within DEGs/WGCNA module genes.

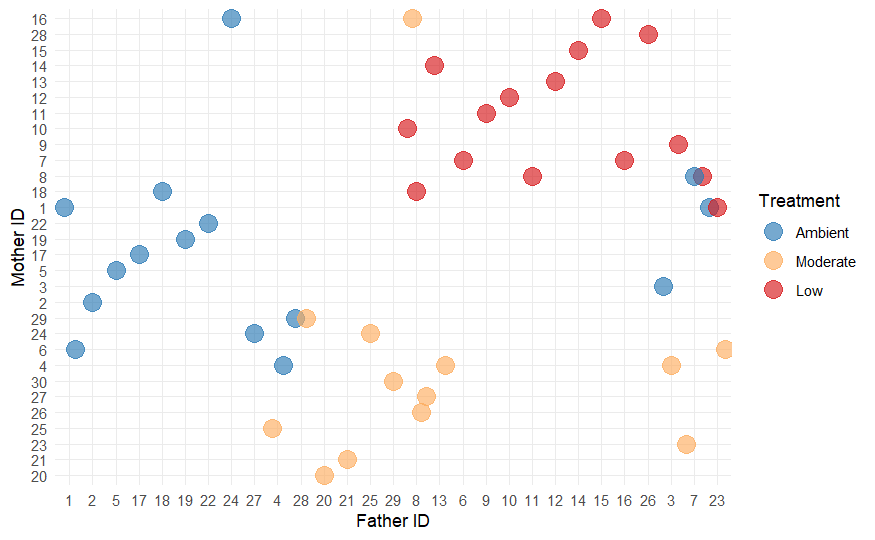




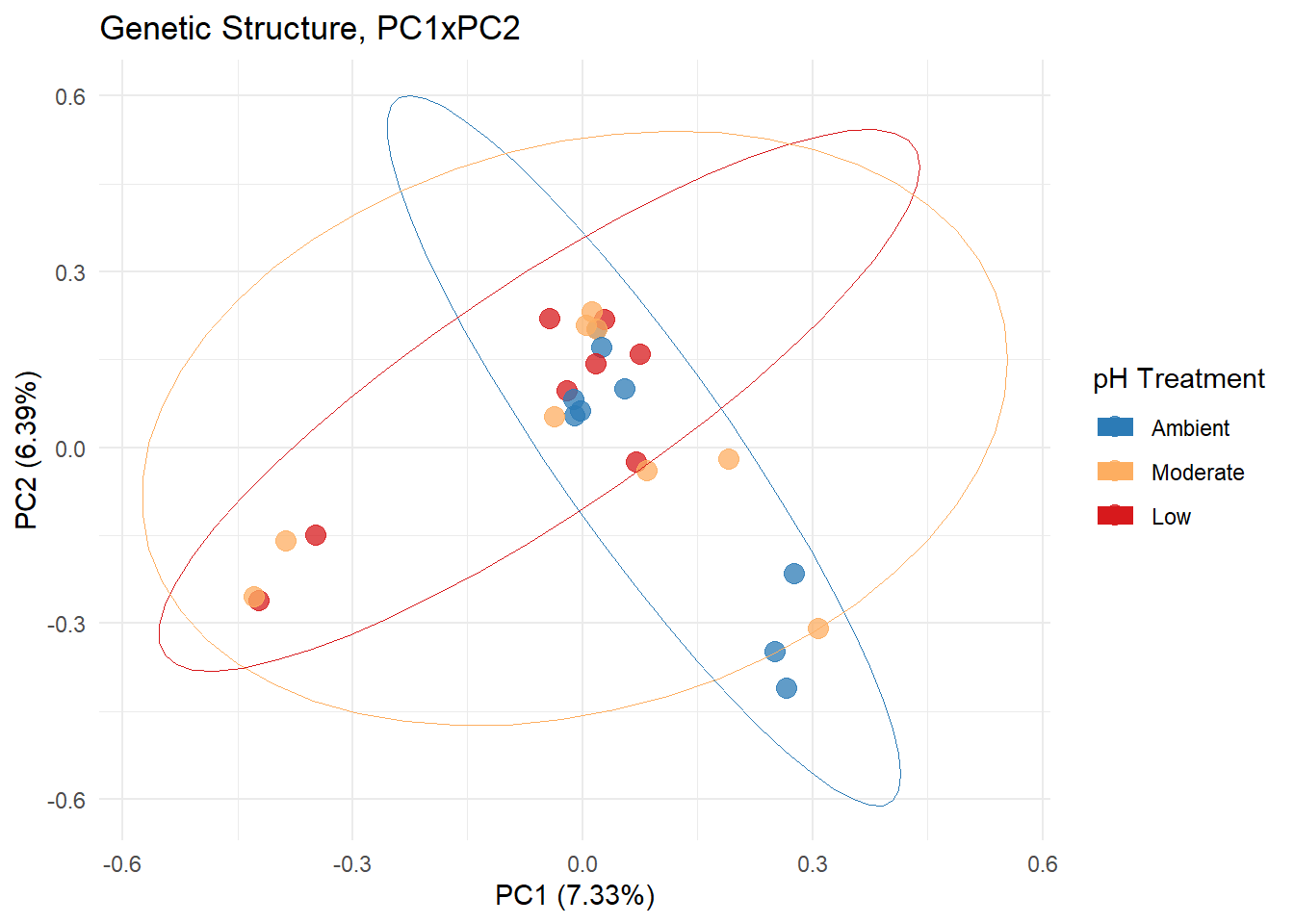
Results from Colony - parentage analysis



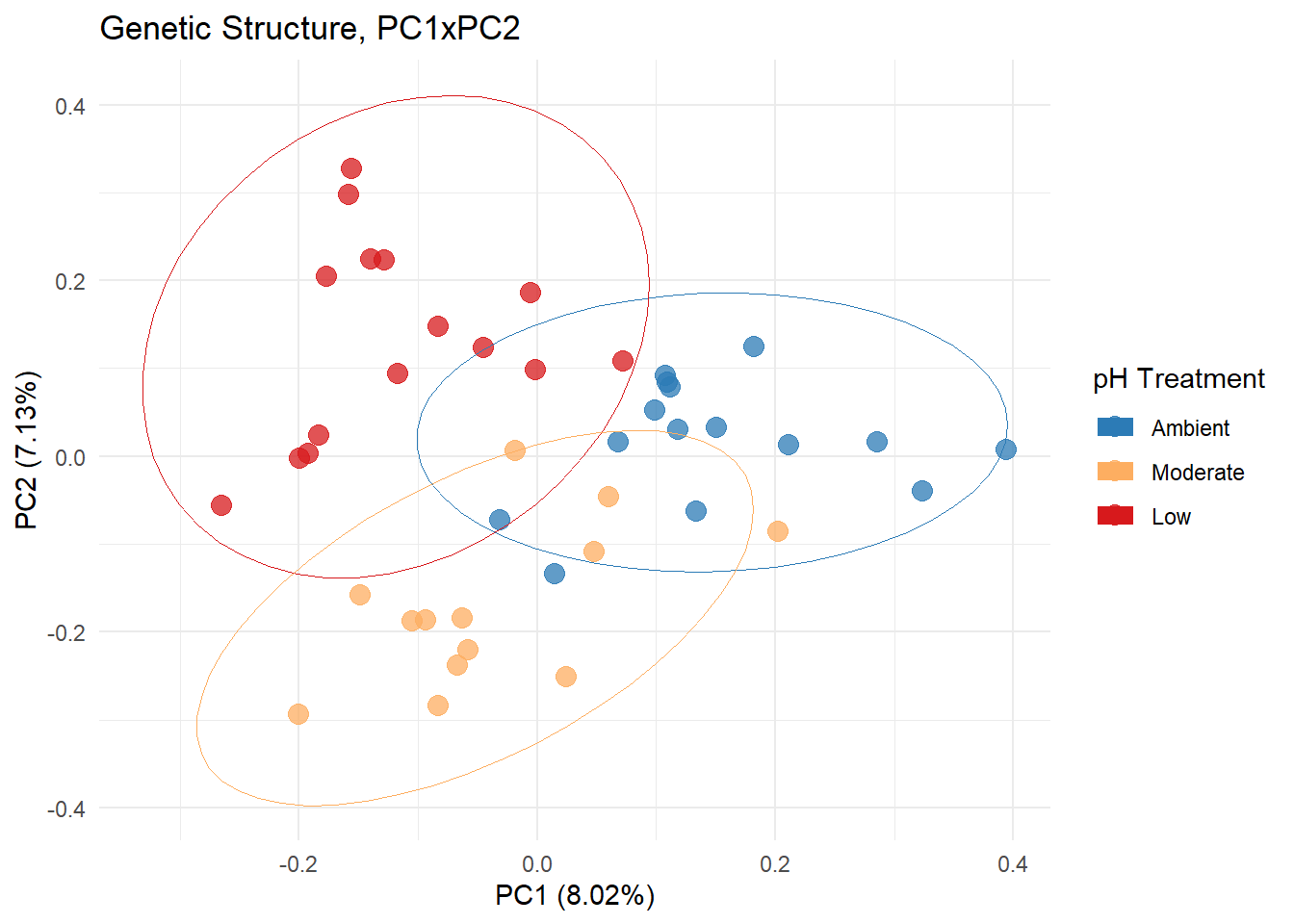




Re-do PCA after removing siblings (half and full), 8-9 samples remain / treatment

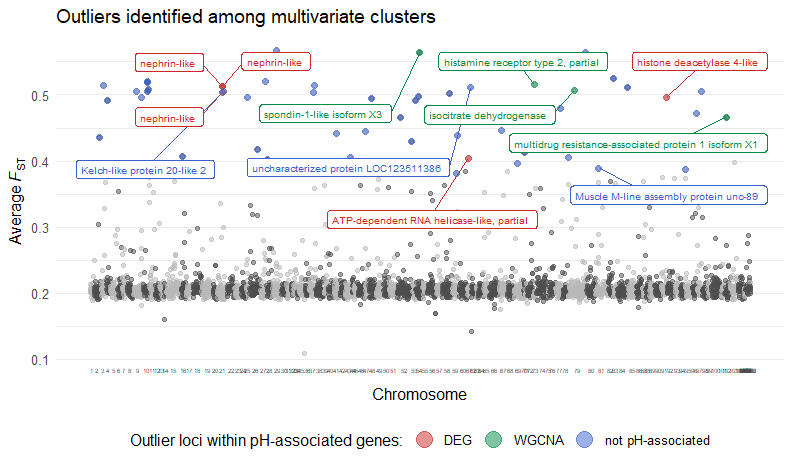


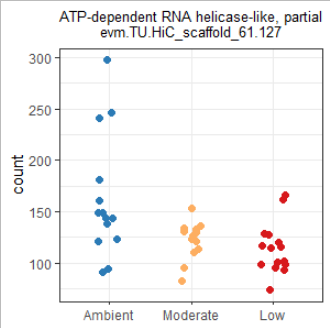
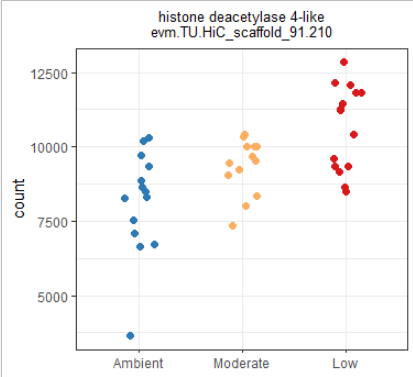
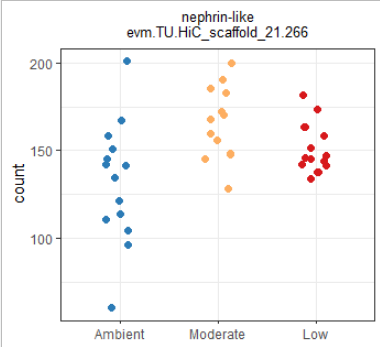
Re-do PCA after only using high Fst values



Results from SNP outlier analyses:

* PCAdapt: 229 outlier loci
* OutFLANK: 0 outlier loci
* Bayescan: 23 outlier loci among clusters (none among pH treatments) - shown in figure (note: histone deacetylase & ATP-dependent RNA helicase are also in WGNCA modules)





## Apr 19, 2022

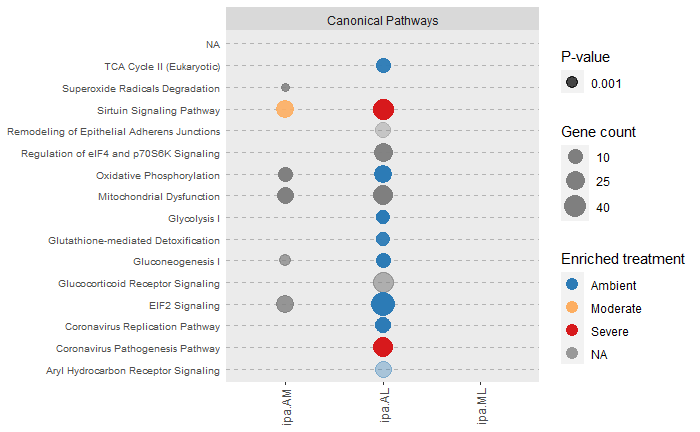
* Pulling SNPs from RNSeq data is often done using STAR, so I’ve re-started the genotyping pipeline using the STAR-aligned .bam file.
* With this in mind, if I want to use any of the STAR-id’d SNPs I’ll also want to use STAR to do the alignment for my RNASeq data. The problem that I encountered before was that STAR’s internal gene quantification step throws out all multi-mapped reads. Using featureCounts I can quantify gene counts from the STAR-based bam files, but the only options are to a) count all multimapped reads, and b) count them fractionally. Neither is my preferred method (in contrast, Bowtie2 identities the best location to map each multimapped reads). Using a model-based quantifier is a good approach, like RSEM, which takes fragment/gene length into account. So, I’d like to use STAR+RSEM.
* To use STAR+RSEM, I need to go back to the alignment step, and adjust the STAR code to output alignments using transcriptome coordinates, which can then be fed into RSEM. E.g. see [this pipeline](https://ycl6.gitbook.io/rna-seq-data-analysis/rna-seq_analysis_workflow/mapping_with_star).

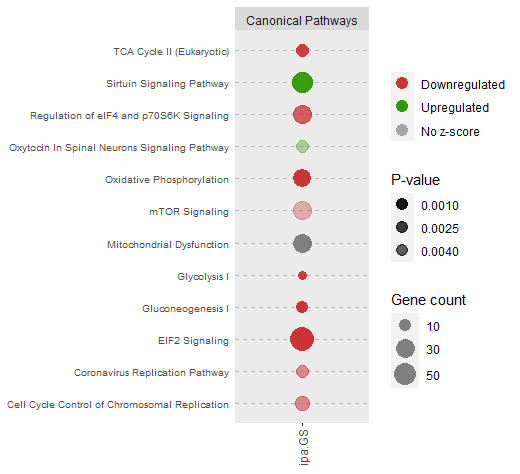
### May 9, 2022

IPA results

We ran IPA on the differentially expressed gene sets, and on the WGCNA modules. For each, we used p-value to designate significance of each gene, and the background set of genes used for enrichment included all genes in my dataset that mapped to IPA genes. Here are two figures showing the top canonical pathways that were enriched, using a maximum p-value=0.0025 for visualization purposes.

This first one shows the top canonical pathways in each of the differentially expressed gene sets.





Notes from meeting with Chris

* OA oxygen consumption study
* Look into urchin species
* Cell signaling - regulating pH using neurotransmitter levels, which may malfunction in low pH. Maybe need more.
* Binding
* More energy in preventative immune response in ambient group
  + In tanner crab - had similar levels of hemocytes, but many more dead hemocytes in OA so higher rate of production
* Lots more energy into moving ions in and out to maintain homeostasis
* Slow metabolism down to produce less CO2 yourself
* Transcriptional regulation
* They were just 5 days post molt - 5 days after 100% of them had molted. Definitely intermolt, since that last 3-weeks to 1 month.
  + They molt
  + Cuticle is formed, but have to calcify that cuticle.
  + In adults, the cuticle hardness takes 120 days to take final cuticle hardness levels after molt. So it’s a process that takes a long time
* Growing faster, require more active individuals in ambient, so need to do all those things faster
* Chris’ data - no real effects to larvae
* The juveniles here - no real phenotypic data - didn’t get good data since growth was poor in ambient conditions so no phenotypic data to go along with this. N
* They were reared from ~3-8C, while
* Can I look at alternative splicing
* Shift in energetics away from growth, towards homeostasis processes. That explains why we’re seeing the transcriptional response -
* Metabolic shrinkage or tightengin

Why not a bigger response between 7.8 and 7.5?

* At 7.8
* At 7.5 they’ve maxed out their options

Larvae are well adapted / capable -

Which journal? Think about that.

## Jun 3, 2022

Revisited the appropriate reference to use for alignment. This is because I lose ~70% of my aligned reads at the gene quantification step. I thought that perhaps using the red king crab transcriptome generated by [Stillman et al. 2020](https://doi.org/10.1017/S002531541900119X) would be a better reference than the blue king crab genome. However, while seeking out the transcriptome file on NCBI I came across a RKC genome! It was published in [Veldsman et al. 2021](https://doi.org/10.1186/s12864-021-07636-9), and it used the Stillman transcriptome data, and included an annotation file (.gff) identifying locations of genes, exons, etc. Data files are published here: <https://doi.org/10.5281/zenodo.4589425>. So, I am re-aligning the data using both Bowtie2 and STAR. I am also extracting gene sequences (in fasta format) from the genome, and blasting those against uniprot/swissprot. What’s my end goal? I don’t need the alignment rate to be perfect, or even as good as alignment to the blue king crab genome (which was ~78% on average), but I’d like the total number of reads aligning to genes (i.e. gene counts) to be more than what I currently have.

* [Script for Bowtie2](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/align_bowtie2_RKC-genome.sh) alignment using Paralithodes.camtschaticus genome
* [Script for STAR](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/align-STAR_RKC-genome.sh) alignment using Paralithodes.camtschaticus genome. *Note: this didn’t complete as the job ran out of memory!*
* Script for blasting the Paralithodes.camtschaticus genome against uniprot/swissprot

Results of RCK genome alignment using Bowtie2:

* Great alignment rate
* Much higher read assignment rate!

## Jun 6, 2022

Can’t run WGCNA on my laptop with the new gene count matrix that I generated using the Red king crab genome + annotations during alignment. Ran WGCNA on Sedna. This can be done via a slurm script with R code. However, it’s also possible to run R code interactively on Sedna. To do this, I initiated an interactive job on a higher memory node via ` srun --pty --mem=180GB /bin/bash`, loaded the R software via `module load R/4.0.3` then started R via `R`. This allowed me to use the R language on Sedna via my terminal to test out code.

## Jun 15, 2022

*Differential transcript analysis.*  Initiated a job to identify isoforms/transcripts using the script [stringtie\_isoforms.sh](https://raw.githubusercontent.com/laurahspencer/red-king_RNASeq-2022/main/scripts/stringtie_isoforms.sh), modified from [Sam’s script](https://github.com/RobertsLab/sams-notebook/blob/master/sbatch_scripts/20220225_cvir_stringtie_GCF_002022765.2_isoforms.sh) described in his [notebook post](https://robertslab.github.io/sams-notebook/2022/02/25/Transcript-Identification-and-Alignments-C.virginica-RNAseq-with-NCBI-Genome-GCF_002022765.2-Using-Hisat2-and-Stringtie-on-Mox.html). I’m following his workflow, described [here](https://robertslab.github.io/sams-notebook/2022/04/25/Project-Summary-C.virginica-CEABIGR-Female-vs.-Male-Gonad-Exposed-to-OA.html). Once Stringtie is finished I’ll use Ballgown to identify differentially expressed transcripts. My goal here is to determine whether any processes/pathways/functions that aren’t differentially expressed at the gene level do in fact differ in terms of splicing. I am curious about this, because I see a lot of transposable element activity, transcription regulation, and some demethylase activity in my DEG analysis. So, my theory is that while I see a general suppression of transcription in the OA treatments, perhaps there are some unique transcripts being produced by OA-exposed crab, regulated by demethylation in gene bodies and due to TE activity. If so, this could be due to a) some type of adaptive mechanism resulting in new phenotypes, or b) everything going awry - less of control of the mechanisms that typically control splicing.