EG homework 2

Your assignment here is to **focus exclusively on the Day10 samples** to address the biological questions that we initially laid out at the start of this section. Namely:

a. Is there evidence that trees from different source climates differ in gene expression?
b. Is there a gene expression response to heat stress? Does this change when the additional stress of drought is added?
c. Is there evidence of an interaction between source climate and stress treatment, such that families from hot/dry climates have a unique expression response to heat or heat+drought compared to families from cool/wet climates?
d. Which genes show the greatest biological response to one of the comparisons above, and what is known about their functional annotation (from congenie.org)?

For each question, you should base your inference on the # of significant genes for the comparison of interest. You'll have to think carefully about how to set up your design and your extraction of results in DESeq to get the comparison of interest that you need for each question.

- The write-up should be 2 pages (max) single spaced, including tables/figures (references on a separate 3rd page)
- You may use a max of 3 tables/figures total. Be sure each table/figure has a title, and a very brief legend describing its contents.

https://rstudio-pubs-static.s3.amazonaws.com/329027_593046fb6d7a427da6b2c538caf601e1.html

Background (1-2 paragraphs):

- A brief description providing context and motivation of the problem we're trying to address with these data
- Brief background on the study species, biological samples, library prep, and sequencing strategy (look through early tutorials for info, plus your notes from class)

Red spruce inhabits cool moist habitats in the northeast. occationally experiences warmer and drier conditions

sampled pops from diff climates - transcriptomic response to heat or heat+draught

Ten maternal families total, 5 from **HotDry, 5 from CoolWet**

Treatment:

- **Control, C:** watered every day, 16:8 L:D photoperiod at 23C:17C temps
- **Heat, H:** 16:8 L:D photoperiod at 35C:26C temps (50% increase in day and night temps over controls)
- **Heat+Drought, D:** Heat plus complete water witholding

Extracted RNA from whole seedlings (root, stem, needle tissue)
Aimed for 5 biological reps per Trt x SourceClim x Day combo, but day5 had few RNA extractions that worked

- Samples were quantified for RNA concentration and quality on the Bioanalyzer
- Samples >1 ng/ul were sent to Cornell for 3' tag sequencing
- Library prep followed the LexoGen protocol and sequencing was on 1 lane of a NextSeq500 (1x86 bp reads)
- Samples were demultiplexed and named according to the convention: POP_FAM_TRT_DAY

Bioinformatics Pipeline (2-3 paragraphs):

- Detailed description of the various steps you used for the analysis of the sequencing data. Take it from QC assessment of the raw reads up to estimation of differential gene expression.
- This section should demonstrate both your technical knowledge of the flow of the different steps in the pipeline, and your level of proficiency in understanding why each step was done. Include justification for using particular analysis approaches or choices as appropriate (e.g., How did we decide what we should use for a reference transcriptome?; In DESeq2, did you filter out genes that had fewer than 1 transcript per individual (on average) and why? Or perhaps you decided to implement a more or less stringent filter -- explain if so.).

Cleaning
FastQC on raw reads –> Trimmomatic –> FastQC on cleaned reads
Reference transcriptome:
`/data/project_data/RS_RNASeq/ReferenceTranscriptome/Pabies1.0-all-cds.fna.gz`
Downloaded from Congenie.org

Before trimming the average number of sequences per sample was 5,007,268.2, which decreased to 4,794,817.8 based on FastQC analysis. Trimmed reads where then used for mapping (see Table 1).

Trimming parameters:
"First, adapter and other illumina-specific sequences were cut from the reads, then bases below the threshold quality = 20 were cut off from the start and the end of a read. Finally, reads were scanned with a 6-base wide sliding window and were cut when the average quality per base dropped below 20. Trimmed reads shorter than 35 were dropped." - did the same here but one

more parameter: HEADCROP: Cut the specified number of bases from the start of the read, here 12 - probs because it was 3' RNA seq

(for population ASC)

Before trimming	After trimming
4939953	4720960
4588368	4423900
5276445	5060372
5494648	5267659
4736927	4501198
5007268.2	4794817.8

Use Salmon to quantify transcript abundance
Use Salmon to simulateously map reads and quantify abundance.

First, index reference "salmon index" - what is this?
-k flag sets the minimum acceptable length for a valid match between query (read) and the reference. Here: 31 (how long were the sequences to begin with?)

Second,
basic command for running the quantification from the documentation, Salmon tutorial

salmon quant
--validateMappings: Enables selective alignment of the sequencing reads when mapping them to the transcriptome. This can improve both the sensitivity and specificity of mapping and, as a result, can improve quantification accuracy.
library type "A"

Mine:
28.5599%, 41.9786%, 32.8468%, 22.5165%, 36.1068%
ASC_06_C_0, ASC_06_C_10, ASC_06_C_5, ASC_06_D_H_0, ASC_06_D_H_10
Percent of reads that mapped to reference

Last time we mapped to only exomes -> low quality. If we include 3'UTR as well -> our mapping rate improved dramatically, ranging from 40-70% of reads mapping across samples, mean of 52%.

What is our mapping rate? Is this good/enough? What factors could affect mapping rate?

DESeq2

We need 2 things:

1. Counts data matrix

Combine individual `quant.sf` files from their respective directories into a counts data matrix with all 76 samples in one table - using R package `tximport`

2. `RS_samples.txt`

table that associates each of our samples with their conditions (climate, treatment, day, population).

Explain what counts data matrix is
Write a little on DESeq2 - look at Thomas' notes

Import the data into DESeq2 in R for data normalization, visualization, and statistical tests for differential gene expression.

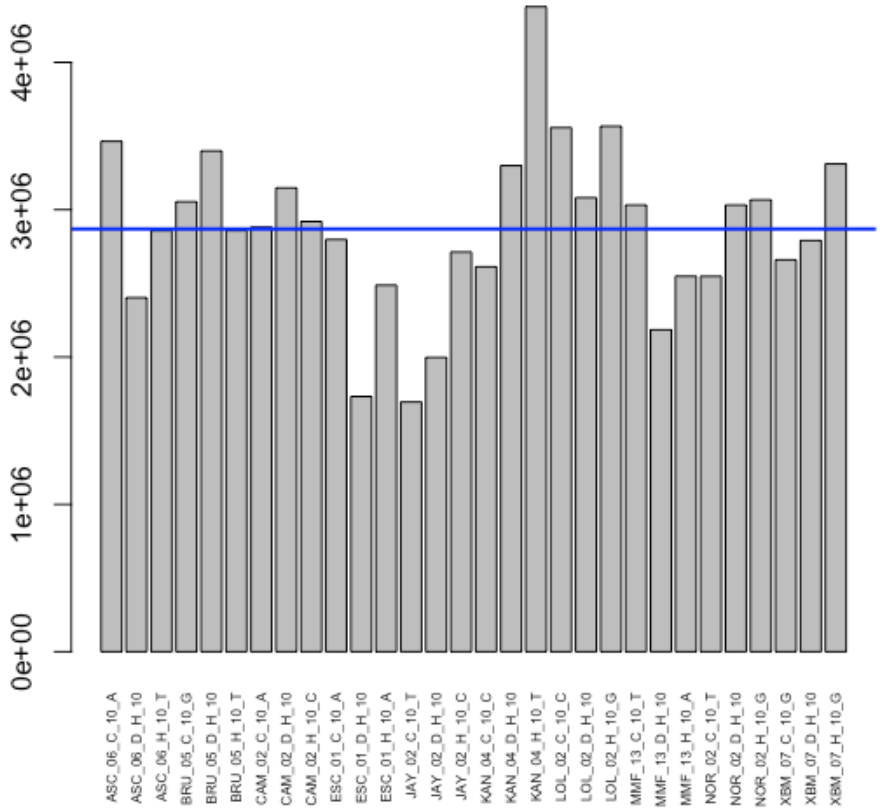
Filter out genes with few reads, if we choose 76 as minimum that is on average 1 read per sample! this reduces the number of tests we have to from 66408 to 23887 -> have to correct less later because of the mulitple stat tests

Results (1-2 paragraphs)

- Report your findings from the different analysis steps. Use a combination of reporting results in-line in your text and summarizing more detailed information in tables and/or figures. You may use a max of 3 tables/figures total. Be sure each table/figure has a title, and a very brief legend describing its contents.
- Include in your results section both "methodological" results (numbers of reads, mapping stats, etc) as well as "biological" results (estimates of log fold change, etc).

From DESeq2
focused exclusively on the Day10 samples

Let's see how many reads we have from each sample - mean, barplot
2869038



What is the average number of counts per gene?
1296.096
Median: 10
not normally distributed gene experssion, some very low/high expression
dispersion across genes - differences in magnitude of expression

What is the average number of counts per gene per sample?
43.2

Options for model: pop, treatment, climate (day is fixed)

Questions:

Results using all data:

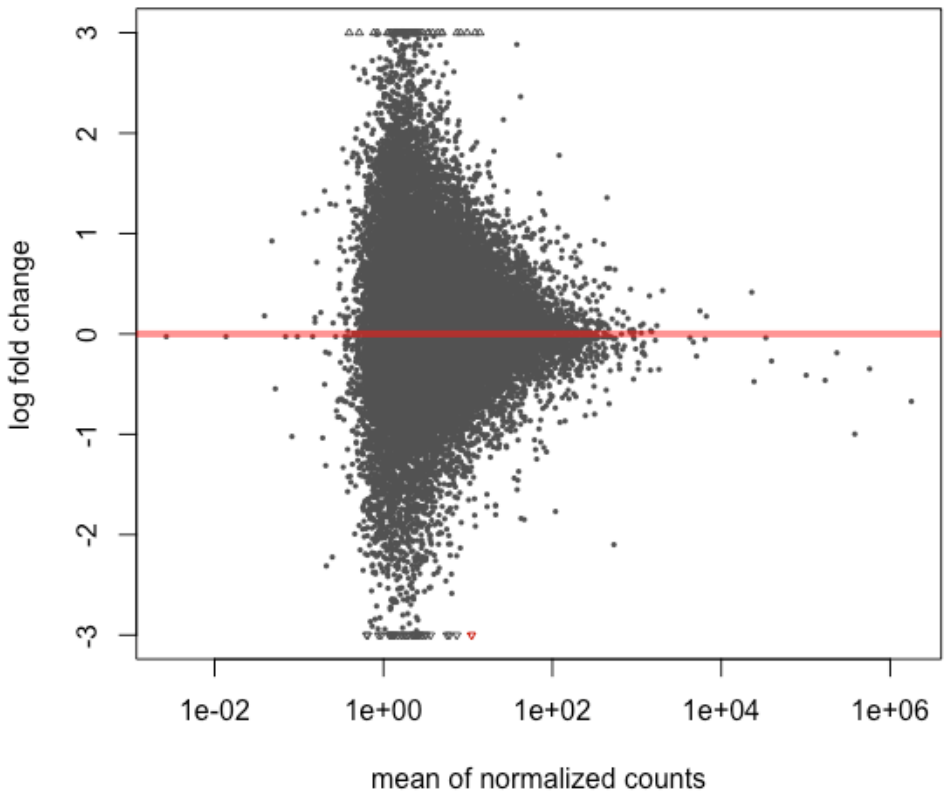
Is there evidence that trees from different source climates differ in gene expression?

Compare climates (HD vs CW) over treatments (C,H,D)
design = `~ climate + treatment + climate:treatment`

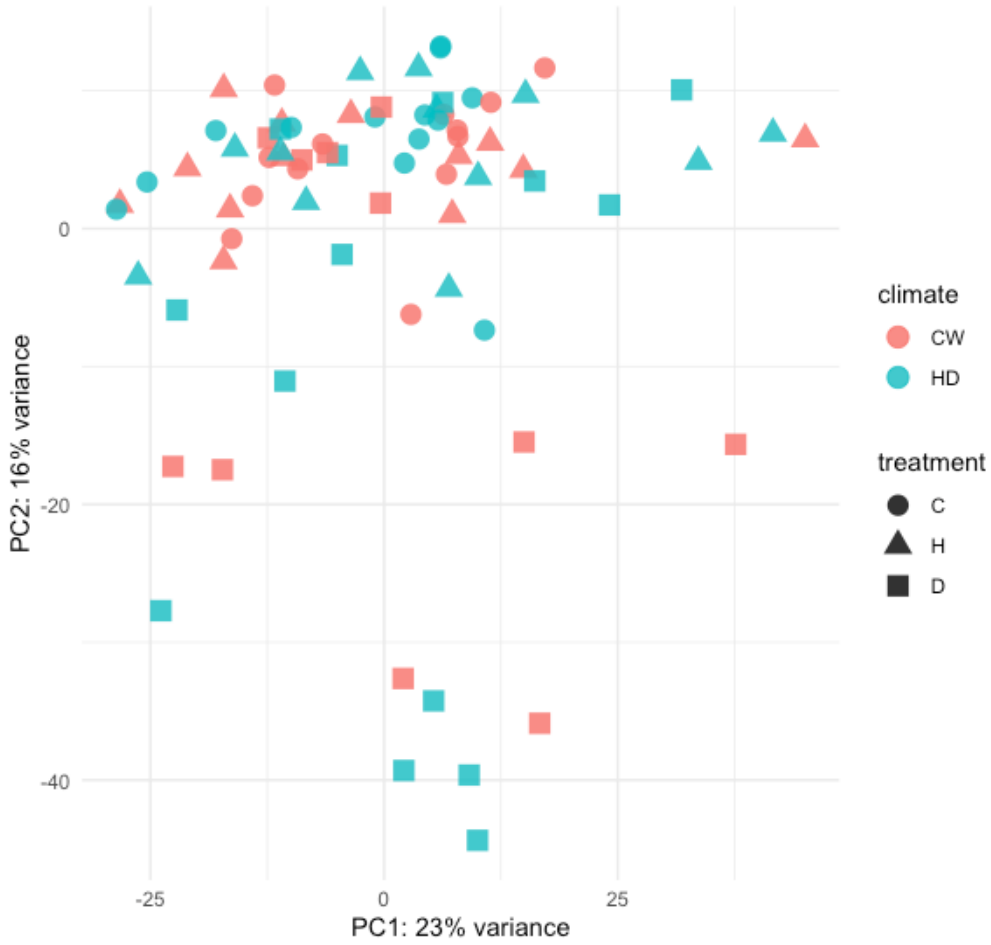
resultsNames(dds)
“Intercept" **“climate_HD_vs_CW"** "treatment_D_vs_C" "treatment_H_vs_C" "climateHD.treatmentD" "climateHD.treatmentH”

climate_HD_vs_CW
across all treatments

```
LFC > 0 (up)       : 0, 0%
LFC < 0 (down)     : 1, 0.0042%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
```



almost nothing is significant, but 1 is significantly and largely dowregulated

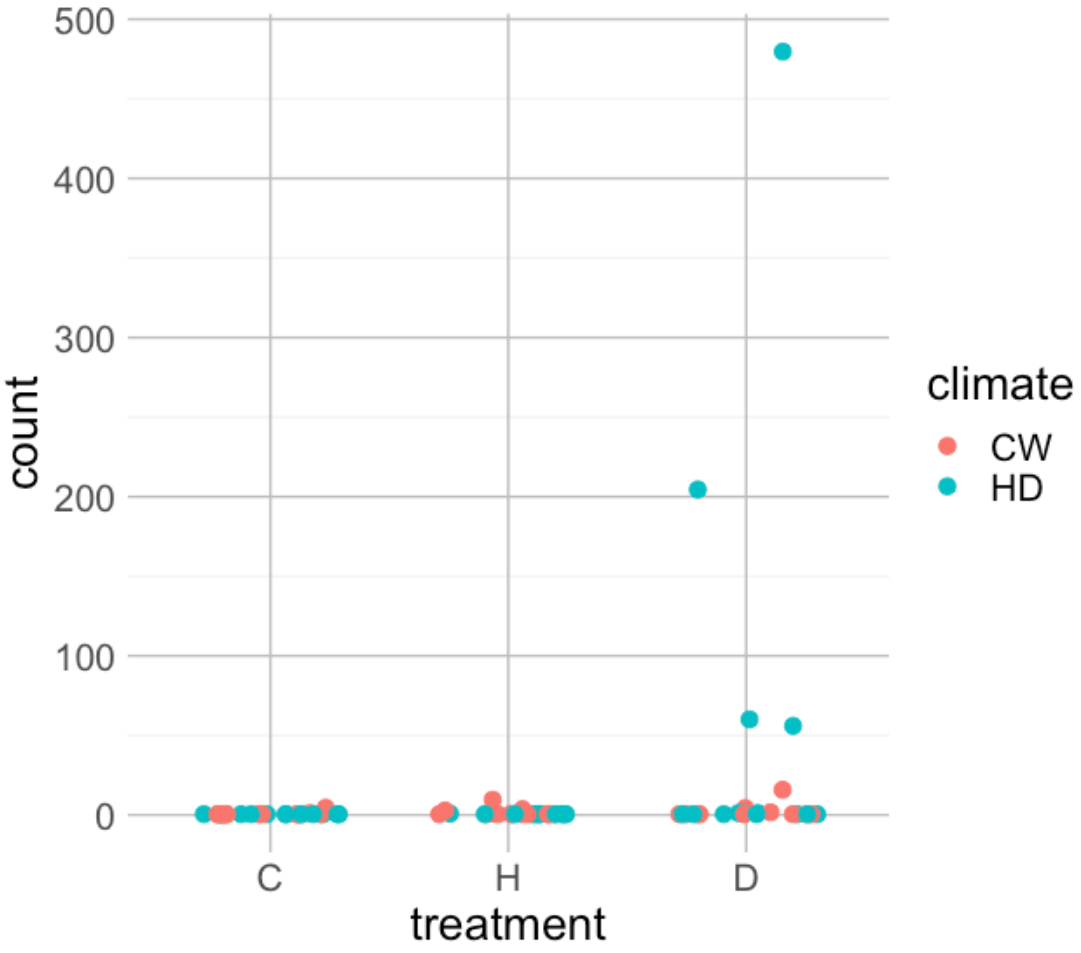


no separation between climates, PC2 separates drought treatment

Counts of specific top gene! (important validatition that the normalization, model is working)
look at a few top genes, if they look similar to what you expect based on model significant result, result it is not driven by one spec gene

There is 1 gene that is significant...
MA_10431378g0010 - **Class VII chitinase, p-value =** 4.97e-10
very much **upregulated in hotdry** | very much downregulated in coldwet, -15.96 logFoldChange

"There are some chitinases, which are expressed in response to environmental stresses, (i.e., high salt concentration, cold, and drought)."
"The accumulation of **chitinase** during **drought** stress was also confirmed by in situ hybridization."
"some of these are pathogenesis related (PR) proteins that are induced as part of systemic acquired resistance."



as you can see upregulated in hotdry, but only in drought! very interesting!

Interactions

Let's look at the interactions now
"climateHD.treatmentD" and "climateHD.treatmentH"

a) the specific effect of treatmentD in climateHD
b) the specific effect of treatmentH in climateHD

a)

```
out of 23887 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 1, 0.0042%
LFC < 0 (down)     : 0, 0%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
```

PCA is the same as above...

we can tell from that : gene experssion was different mainly due to treatment - dry+hot

1 significant:
MA_10431378g0010 - pvalue 1.585e-8, graph as above

b)
none is significant

Awnser: they differ in only 1 gene: indv coming from hotdry conditions upregulate **chitinase in dry+hot conditions**

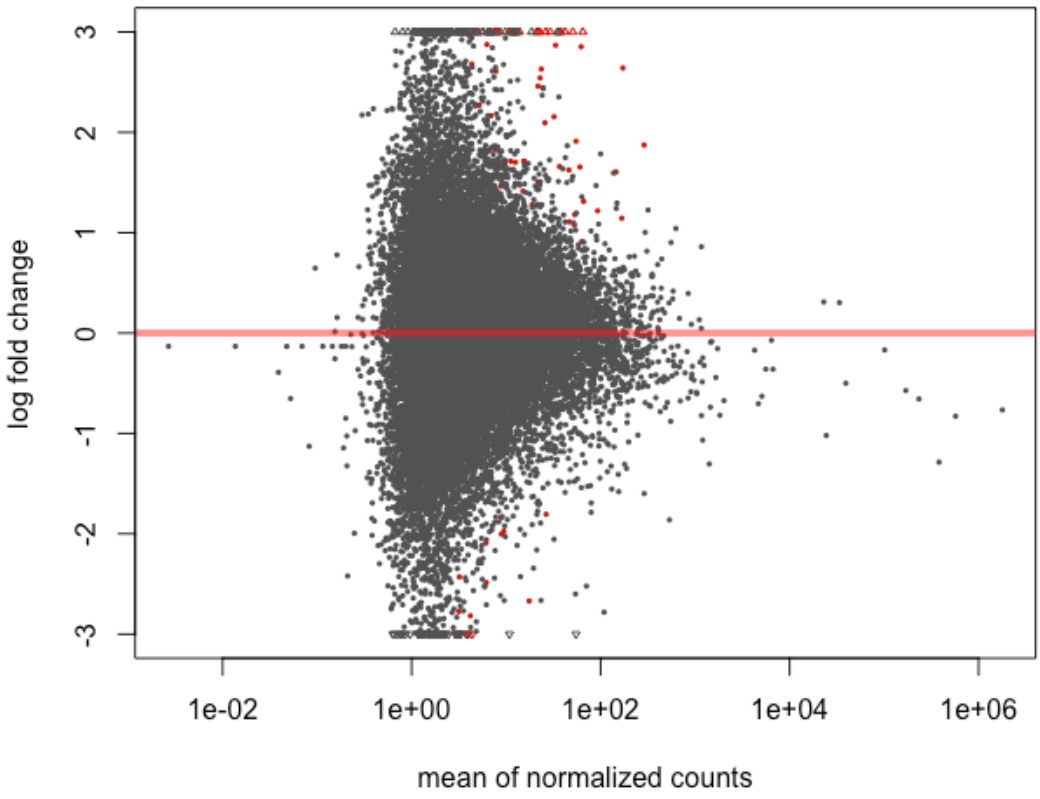
Is there a gene expression response to heat stress? Does this change when the additional stress of drought is added?

This is "treatment_D_vs_C" "treatment_H_vs_C"

"treatment_D_vs_C"
upregulation in drought of 71 genes! downregulation of 16

```
out of 23887 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 71, 0.3%
LFC < 0 (down)     : 16, 0.067%
outliers [1]       : 0, 0%
low counts [2]     : 6947, 29%
```

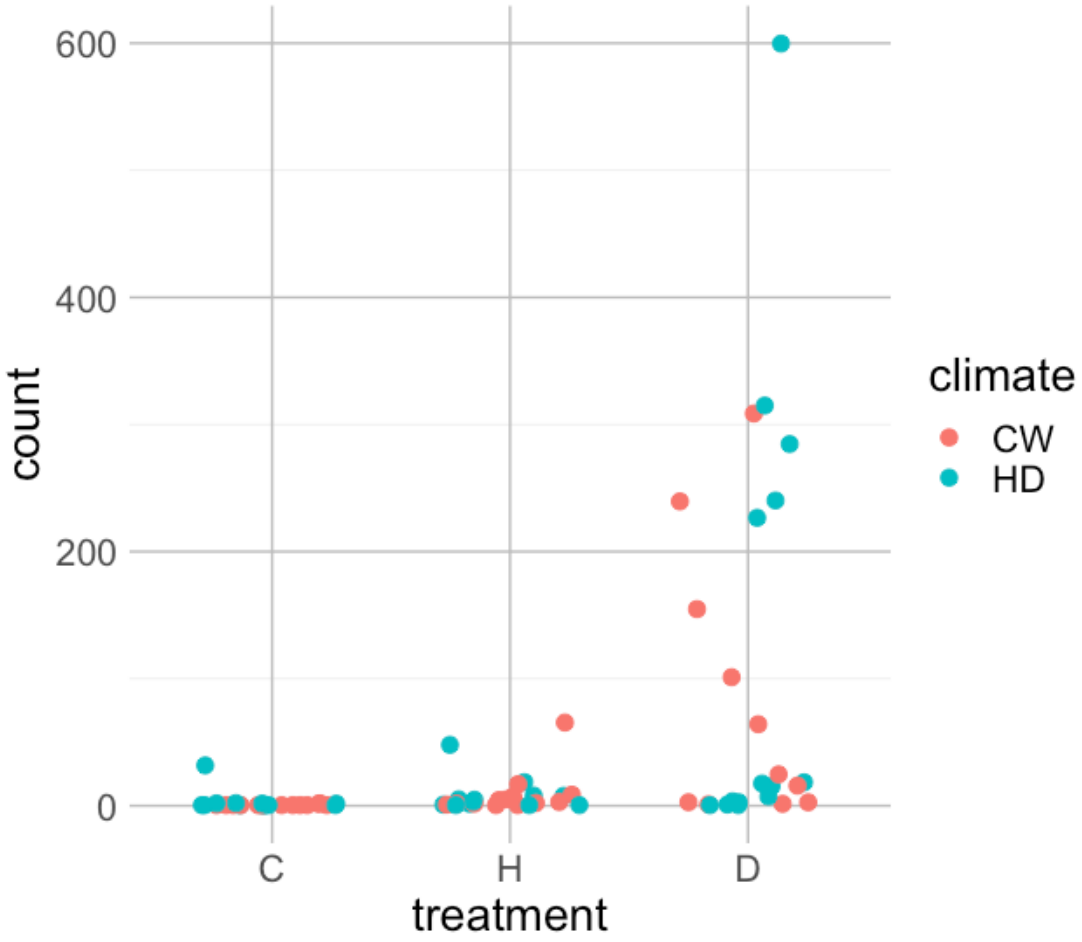
quite a few significant



6 most significant:
MA_75192g0010 4.58334840206522e-07 **NAC domain-containing 68-like**
MA_10257300g0010 8.84670438731186e-06 unknown
MA_7017g0010 0.000332345569183644 unkown
MA_1400g0010 0.000645311548180967 **Late embryogenesis abundant D-34-like**
MA_444738g0020 0.000645311548180967 **Phosphatase 2C**
MA_73034g0010 0.000645311548180967 **Peroxidase**

Gene name	P-value	Up/down	Function
MA_75192g0010	4.58334840206522e-07	positive	
MA_10257300g0010	8.84670438731186e-06		
MA_7017g0010	0.000332345569183644		
MA_1400g0010	0.000645311548180967		
MA_444738g0020	0.000645311548180967		
MA_73034g0010	0.000645311548180967		

MA_75192g0010

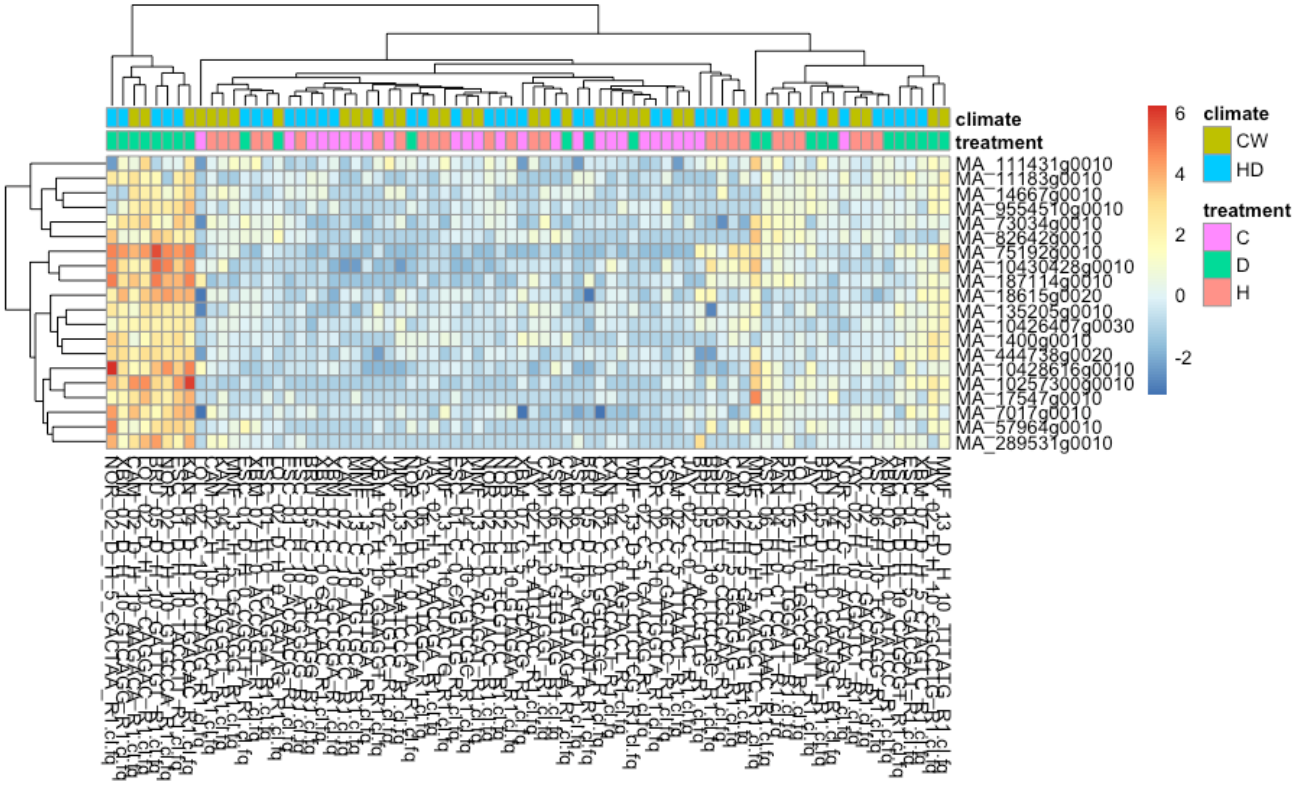


as you can see this is more expressed in drought, but doesn't depend on climate

other 5: very similar graph, MA_444738g0020 looks nice

Again, notice drought has the greatest effect

heatmap:
again, upregulation in drought!!! compared to control



now looking at a few genes downregulated in drought, significant
MA_9068991g0010 unkown
MA_18054g0010 **Probable calcium-binding CML43**
MA_245557g0010 **Auxin-responsive SAUR72-like**

"treatment_H_vs_C"
No significant genes!

a. Is there evidence of an interaction between source climate and stress treatment, such that families from hot/dry climates have a unique expression response to heat or heat+drought compared to families from cool/wet climates? - yes, see above
b. Which genes show the greatest biological response to one of the comparisons above, and what is known about their functional annotation (from congenie.org)? - chitinase

Results using day 10 data:

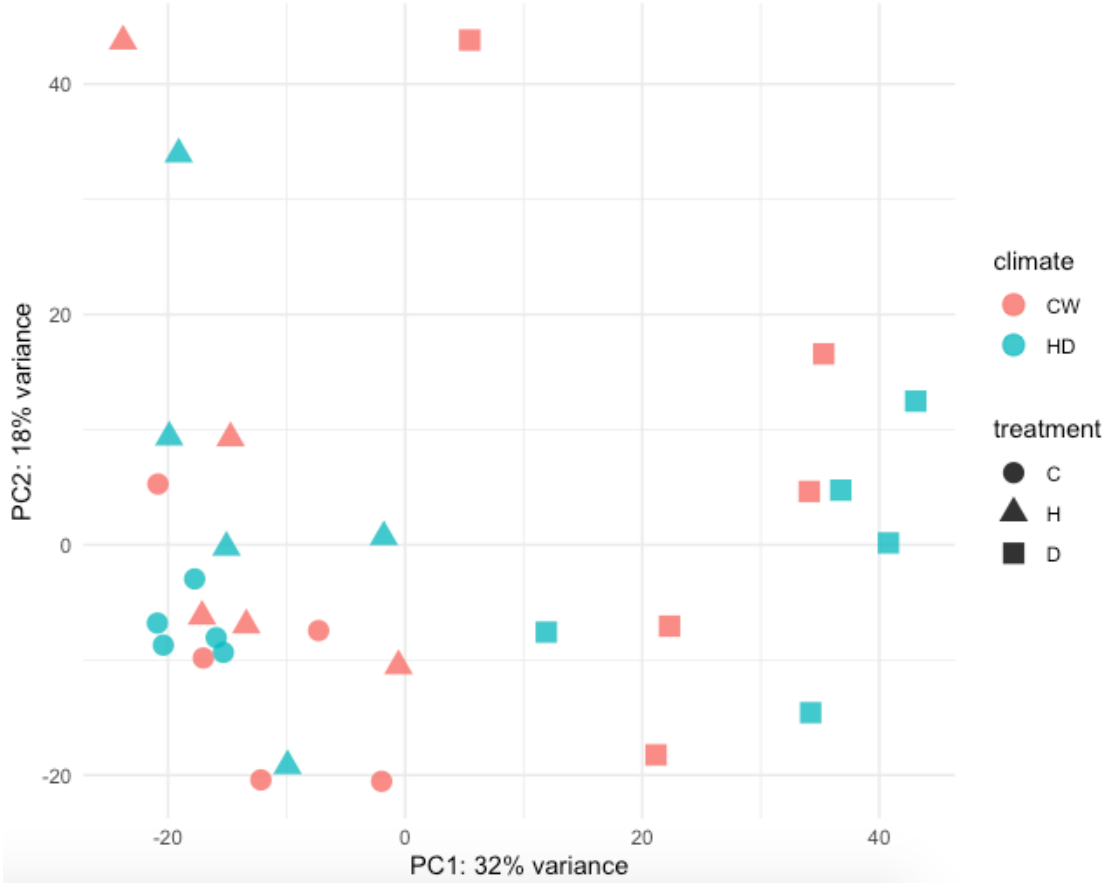
“Intercept" “climate_HD_vs_CW" “treatment_D_vs_C" “treatment_H_vs_C" "climateHD.treatmentD" "climateHD.treatmentH”

climate_HD_vs_CW

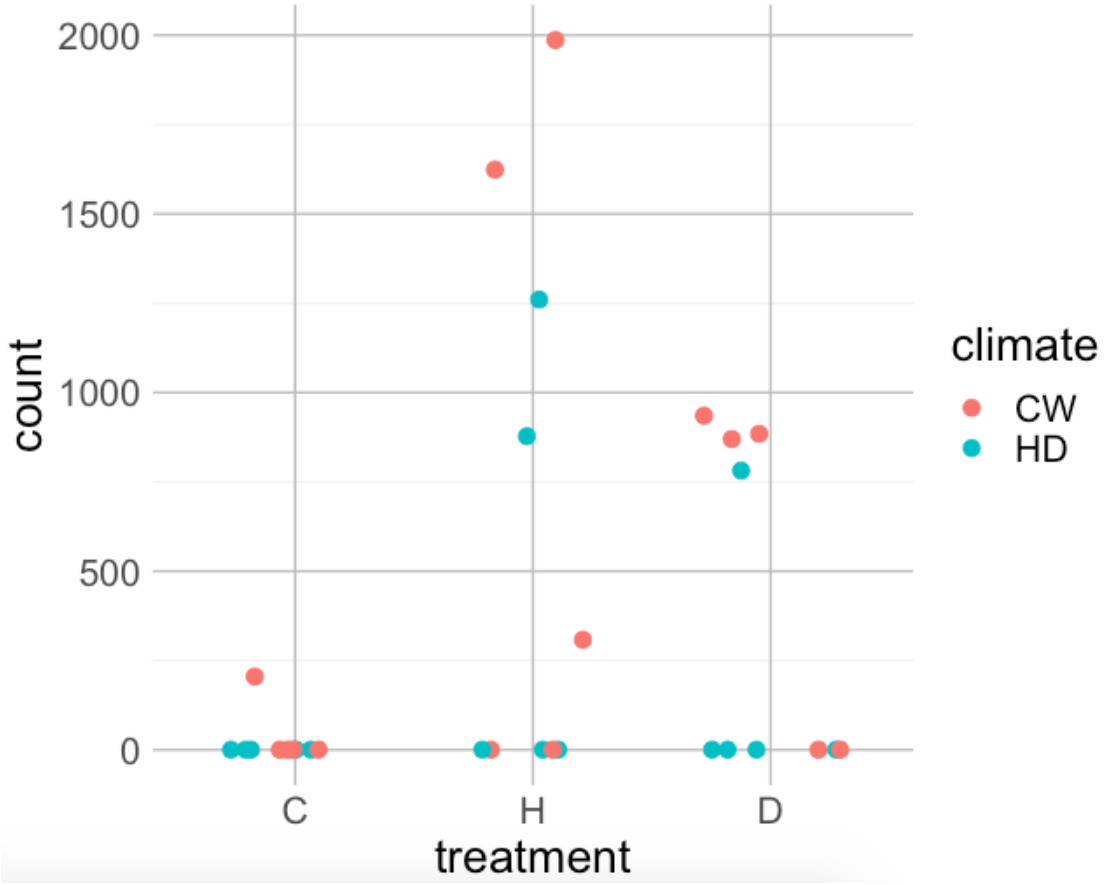
```
out of 17487 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 0, 0%
LFC < 0 (down)     : 1, 0.0057%
outliers [1]       : 38, 0.22%
low counts [2]     : 0, 0%
```

1 gene that is different:
MA_129323g0010, padj: 0.0110269645197478, log2FoldChange: -21.5499067186968
downregulated in hotdry

Ncbi Blast with CDS , Optimize for More dissimilar sequences (discontiguous megablast) still no hit



drought treatment clearly separate, no climate effect



treatment_D_vs_C

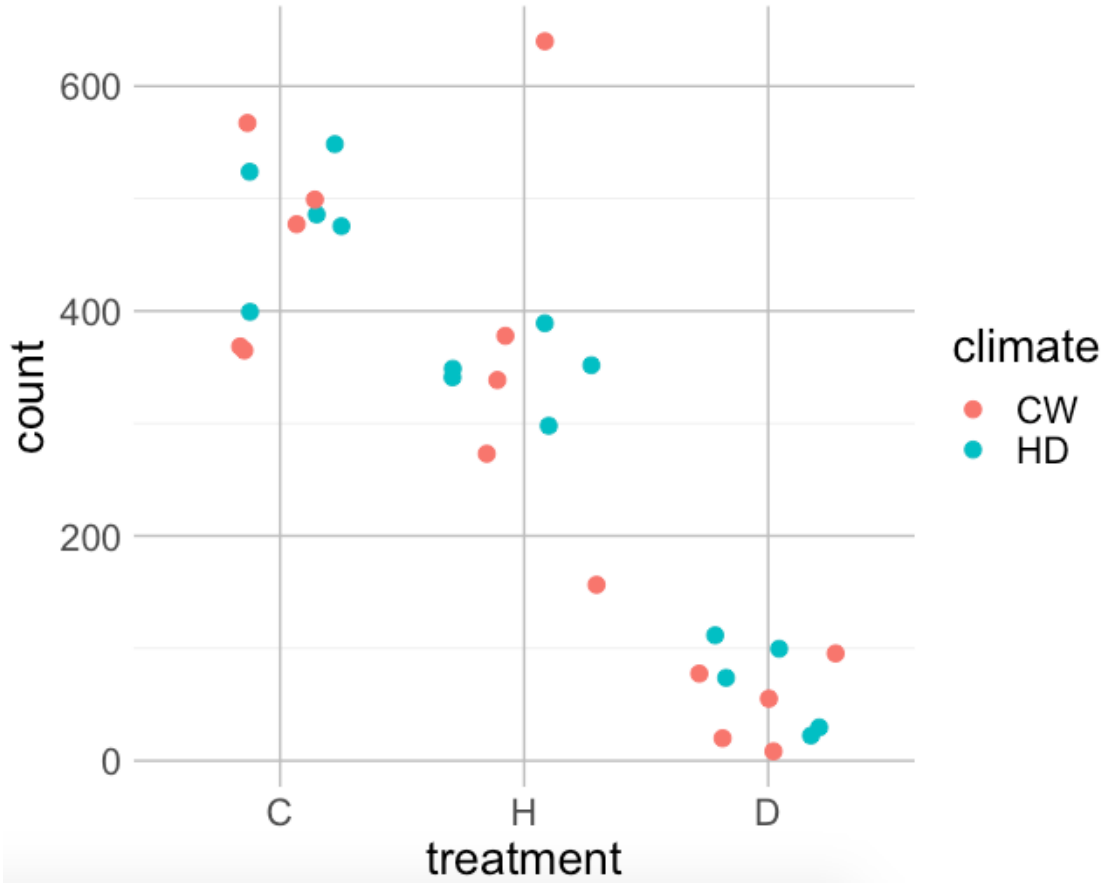
```
LFC > 0 (up)       : 257, 1.5%
LFC < 0 (down)     : 330, 1.9%
outliers [1]       : 38, 0.22%
low counts [2]     : 4043, 23%
```

up = upregulated in drought

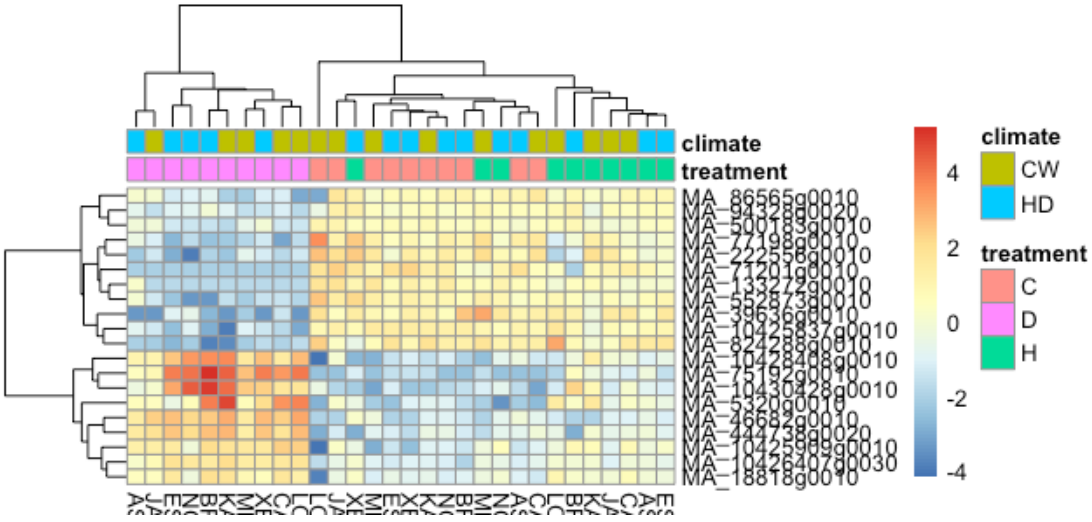
MA_10426407g0030	2.55753580961529	1.14284841387114e-10	**Phox Bem1p**	unknown GO
MA_75192g0010	8.04736203119503	4.84854642258676e-08	**NAC domain-containing 68-like**	regulation of cellular process
MA_10425837g0010	-3.17864866055592	4.84854642258676e-08	NA	unknown GO
MA_824288g0010	-4.15963978724243	4.84854642258676e-08	**Lipid-transfer DIR1**	Probable lipid

				transfer
MA_133272g0010	-6.86530962020971	7.58680909158101e-08	**DNA topoisomerase 2**	DNA topological change
MA_71201g0010	-7.2915108408272	7.58680909158101e-08	**Dirigent**	unknown GO
MA_46682g0010	5.0402868 (8)	1.678388e-07	**Probable phosphatase 2C 8**	unknown GO
MA_10425909g0010	3.1228048 (11)	3.127502e-06	**E3 ubiquitin-ligase ORTHRUS 2**	unknown GO
MA_18818g0010	2.4177939 (12)	4.087774e-06	NA	unknown GO
MA_77198g0010	-4.2290203 (7)	1.236095e-07	**Histone H2B**	nucleosome assembly

1st gene is good for a figure, 3rd is also nice
check all ten above -> ok



this was the 3rd



treatment_H_vs_C

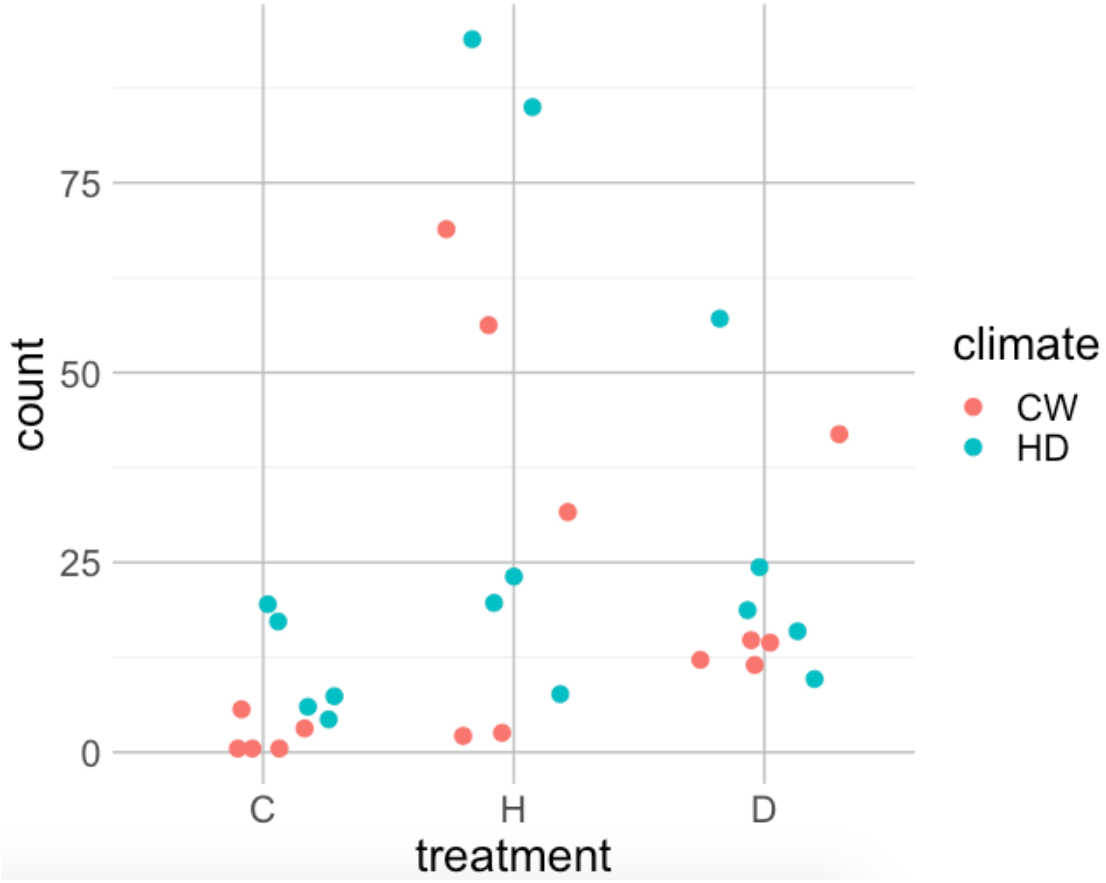
```
LFC > 0 (up)       : 4, 0.023%
LFC < 0 (down)     : 2, 0.011%
outliers [1]       : 38, 0.22%
low counts [2]     : 0, 0%
```

dry had much more effect - again! (like in whole dataset)

MA_10433227g0010**, KDa class I heat shock -like,** 3.26707745791814 (log), 0.00753704414315307, upregulated in heat!!!
MA_10243352g0010, **KDa class IV heat shock -like,** 4.2041318698496 (log), 0.0499479650510409 (adjusetd p), upregulated in heat!!!
MA_10427910g0010**, KDa class I heat shock -like,** 4.44212681158808, 0.0499479650510409

also in drought? 1st not, 2nd not, 3rd yes, upregulated in drought (0.0166190908449393 (apval))

checked all, looks good,



the third heat shock

heatmap doesn't look as good

"climateHD.treatmentD" "climateHD.treatmentH"

nothing significant

Conclusion (1-2 paragraphs)

- Give your biological conclusion so far from the data: What have we learned about the abiotic stressors that elicit gene expression response in red spruce? How does it relate to differences in source climate? What genes are involved? Explicitly use the results to address the 4 questions above.
- Discuss any caveats or uncertainties that should be considered when interpreting the biological conclusions.
- Discuss any methodological challenges encountered along the way that are relevant to your results and their interpretation.
- Discuss opportunities for future directions.

References (listed on a separate page)

- Cite papers in MLA format.
- Github: Have your github lab notebook up to date, and any scripts used for your analysis available in your github "myscripts" folder