EG homework 3

The assignment here is to identify and analyze differentially methylated loci between two different sets of samples. In both cases, we want the control to be AA_F25. This gives us three possible combinations, of which you will **pick two**:

1. AA_F25 and AH_F25
2. AA_F25 and HA_F25
3. AA_F25 and HH_F25

Questions to address:

1. How have methylation patterns changed after 25 generations in different environments?
2. Which environment likely requires the least amount of physiological adjustments as compared to the ambient control (AA)? Which requires the most?
3. What types of genes contain loci that are differentially methylated?
4. What are the similarities and differences in functional targets of methylation between the two treatment conditions?

For each question, you should base your inference on the number of significant epigenetic differences, the degree of overlap between contrasts, and any functional information you can retrieve for the loci.

The write-up should be 2 pages (max) single spaced, including tables/figures (references on a separate 3rd page)

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

Acartia tonsa
worldwide distribution, huge population size, most abundant zooplancton → important ecological role (food chain, biochemical cycle)
large distribution → spans different climates, with varying co2 conc and T

Like many plankton common to estuarine ecosystems, they can live in a wide range of temperatures and salinities.[2]

global warming (ocean acidification, temperature increase)
role of epigenetics in rapid adaptation? prediction and molecular mechanisms

A.tonsa and T:
ideal = 16-17 ºC
Survival was generally high (>70%) at intermediate temperatures (10–20°C) but decreased rapidly above 25°C at all salinities - here

A.tonsa and Co2:
>20% difference was seen in hatching success between experiments at 1000 µatm pCO_2 scenarios (2100 year scenario), and >85% at 6000 µatm pCO_2.
paternal limitation in reproductive success, or a combined maternal and paternal effect
-here

in the conclusion maybe:
3½ yr long selection experiment, *Acartia tonsa* populations in seawater treated with 200 and 800 µatm CO_2
Over the experimental period, beneficial adaptations of the copepods cultured under high CO_2 conditions of elevated seawater pCO_2 and associated food quality were not detected. However, towards the end of the experiment, copepods cultured under elevated pCO_2 and fed with high CO_2 algae showed increased body mass and decreased prosome length. - here

Exp set up
common gardened them for three generations
our treatments with four replicates each and about 3,000-5,000 individuals per replicate
25 generations.

18C for ambient. 22C for high temp. 400 ppm for ambient CO2 and 2000 ppm for high

ambient 0 and 25
high co2 25
high t 25
both 25

reduced representation bisulfite sequencing (RRBS).
Before starting we also spiked in a small amount of DNA from *E. coli* that we know wasn't methylated. Using this, we can calculate downstream how efficient our bisulfite conversion was.

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 methylation.
- 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 you filter out positions by coverage? What methylation difference was required to count a position as significant? and why? Or perhaps you decided to implement a more or less stringent filter -- explain if so.).

Fastqcs look weird - bunch of T, few C, as expected (unmethylated C's were converted to T), + other funkiness \
-> special method of alignment to reference needed \
we need 2 versions of reference: all C's converted to T; all G's converted to A (for other strand), we need to do 2 alignments \

1. Visualise, clean, visualise - fastqc, trimmomatic - was done for us
2. Align to Acartia tonsa reference genome (Bismark)

bismark —bowtie2

also align lambda DNA to check for conversion efficiency

If you remember, we're using two modified versions of the genome where we've converted C-to-T AND G-to-A. We also generate temporary files of our trimmed reads where we convert C-to-T (read1) AND G-to-A (read2) so they can map to this converted genome. This makes the alignment a bit harder because 1. the complexity of DNA is reduced, and 2. we are mapping to two separate genomes.

```
-local align with the local alignment option in bowtie2. This
will include soft clipping, which should increase mapping rat
e, but comes at the cost of (maybe) increasing mis-mapping.
```

3. after mapping we extracted methylation calls with

```
bismark_methylation_extractor
```

4. Process and filter calls

begining of read higher methylation rate, probs error -> we are trimming it off

→ we have now chr, start,stop, methylation percentage in coverage file

5. Test for differential methylation (Methylkit)

red in coverage files, get basic statistics about read coverage per base → # we don't want to consider the very high coverage ones because that migth be due to eg pcr duplicates

```
filter samples by depth
hi.perc=97.5
```

then we calculated percent methylation for each site and sample
did multiple pairwise comparisions (i did AA25 vs AH25 and AA25 vs HA25)
→ future study, better way of doing this?

calculateDiffMeth() = calculate differential methylation between two groups
getMethylDiff() = get the significant ones of these (qvalue=0.05), and the ones that are more than 10% different

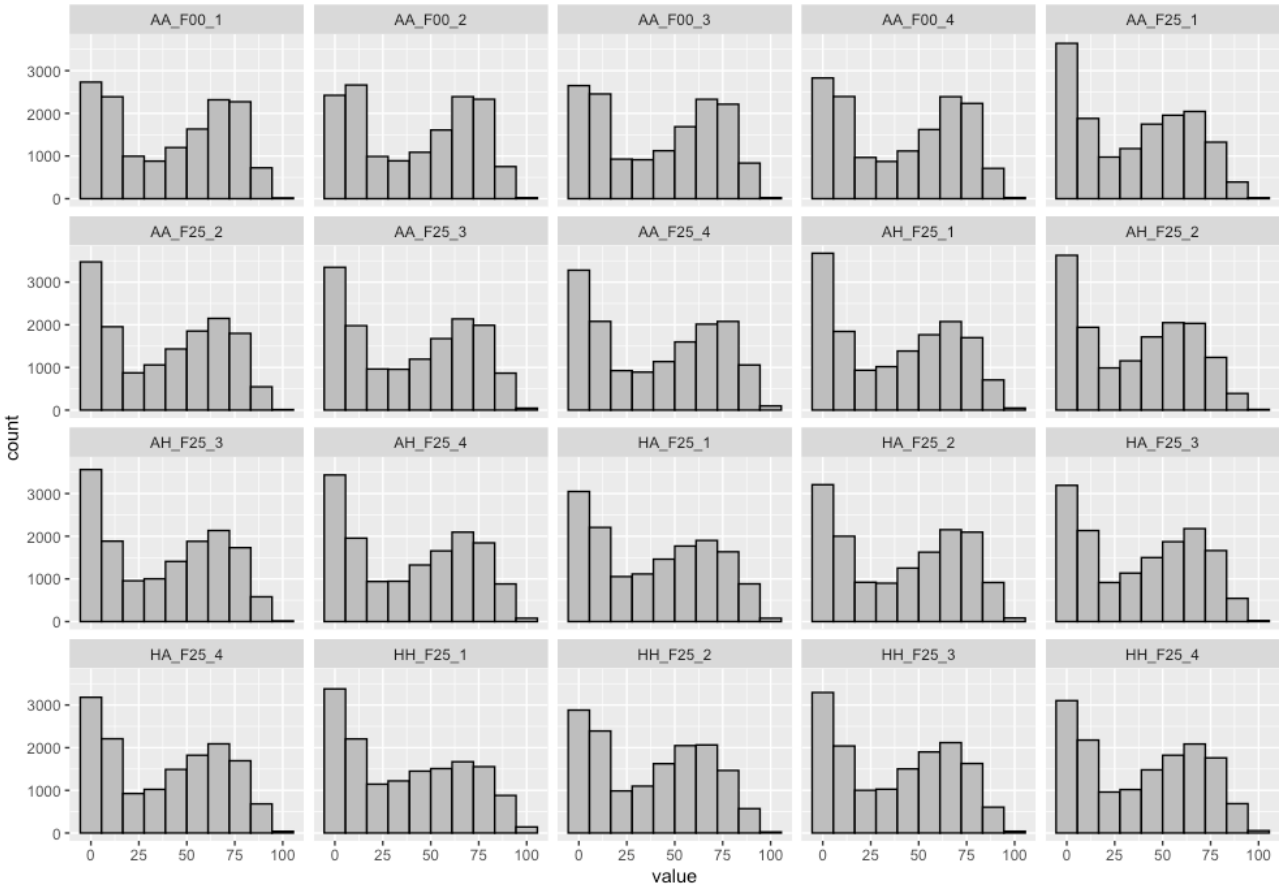
finally used bedtools closest to find SNPs in the annotation table

Results (1-2 paragraphs)

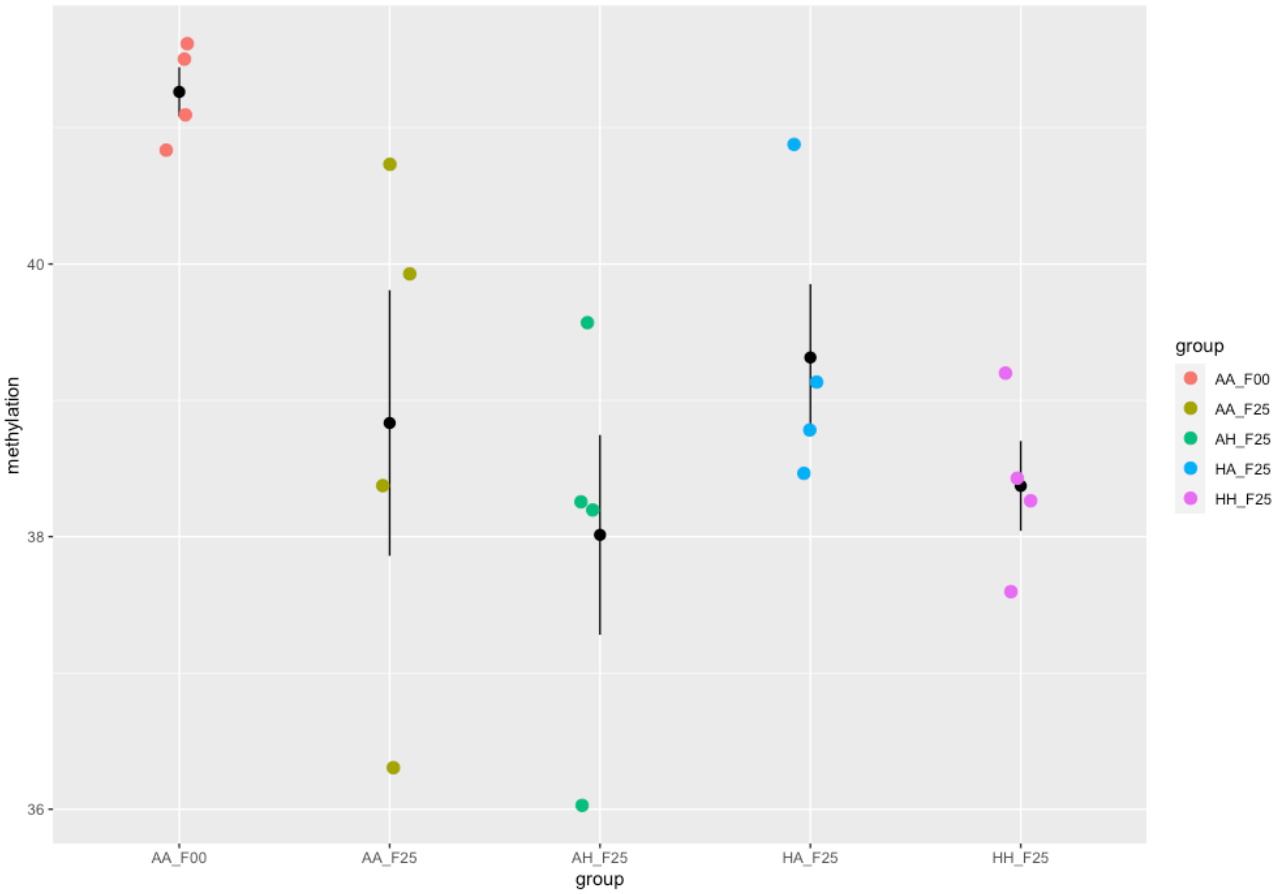
- Report your findings from the different analysis steps. Use a combination of reporting results in-line in your text and summarizing more detailed information in tables and/or figures. **You may use a max of 3 tables/figures total. Be sure each table/figure has a title, and a very brief legend describing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

mapping success: big differences between populations, best AA0, ~65%, the rest ~40%

calculates the percent methylation for each site and sample →
and this is a histogram of that
no major diffs in distribution it seems like



if we take an average for all the samples (ie 20 little histograms) and plot it

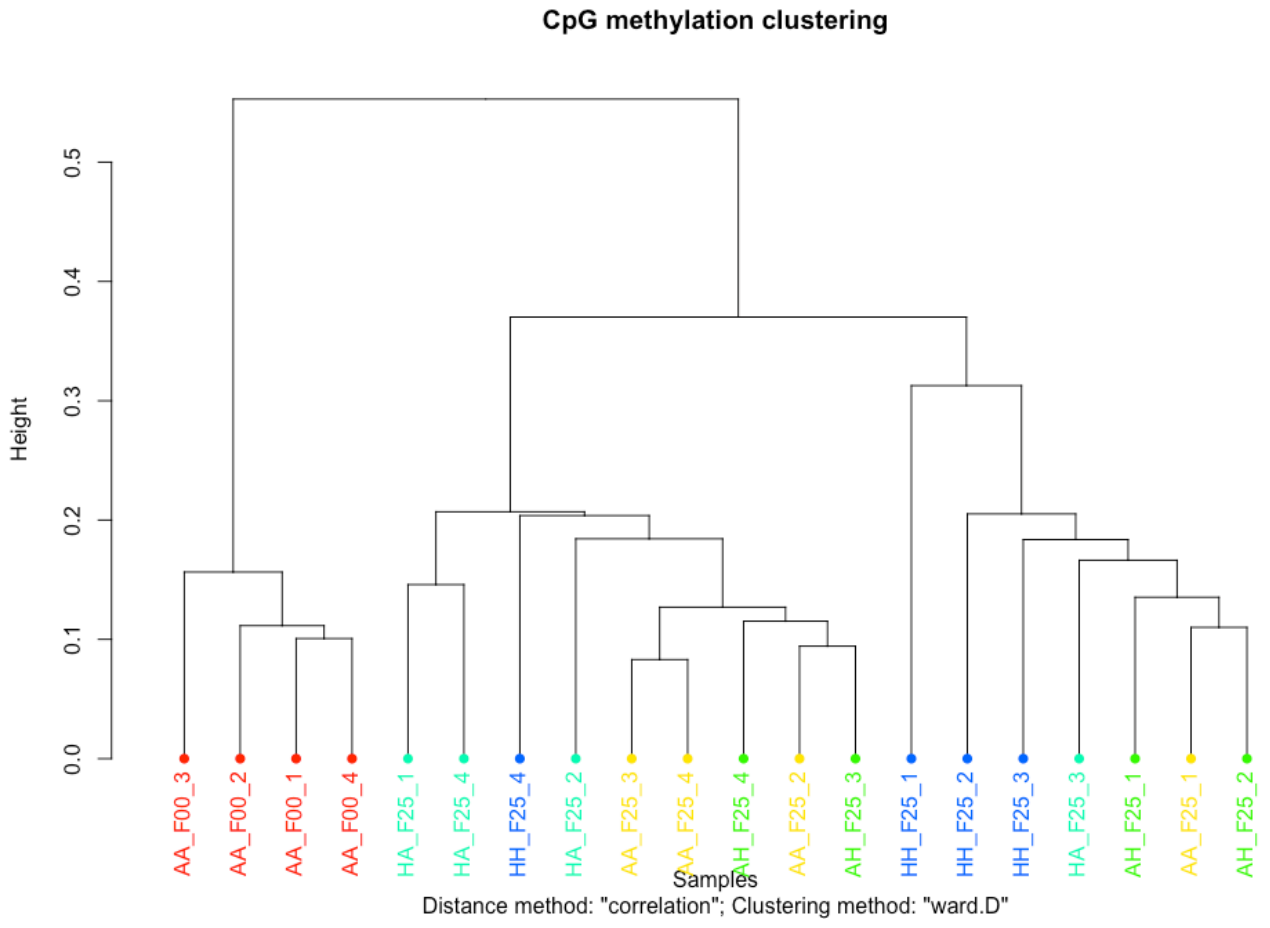


it seems like average frequency of methylation decreased over time, control in time 0 has the highest average methylation. but smaller differences between 25 generation samples
However, AA_F00 was also the one that was more mapped to reference -> difficult to tell we see real diff in average methlyation rate per site here

used ANOVA for above data (comparing 25 gens because those are independent, AA0 is not)

source	sum of squares SS	degrees of freedom ν	mean square MS	F statistic	p-value
treatment	3.8282	3	1.2761	0.6761	0.5832
error	22.6491	12	1.8874		
total	26.4774	15			

interestingly when we use clusterSamples() only AA generation 0 is clustered look this method up and other types of clustering??



→ no changes in average methylation frequency per site
there are also no changes in distribution of frquencies, always most SNPs are not methylated while the rest of the distr follows a bell-curve shape where most SNPs have ~60% methylation rate
but there could be differences in what sites are methylated more or less in diff conditions

find overlapping SNPs!

I think I am going to do the following 2 analysis:

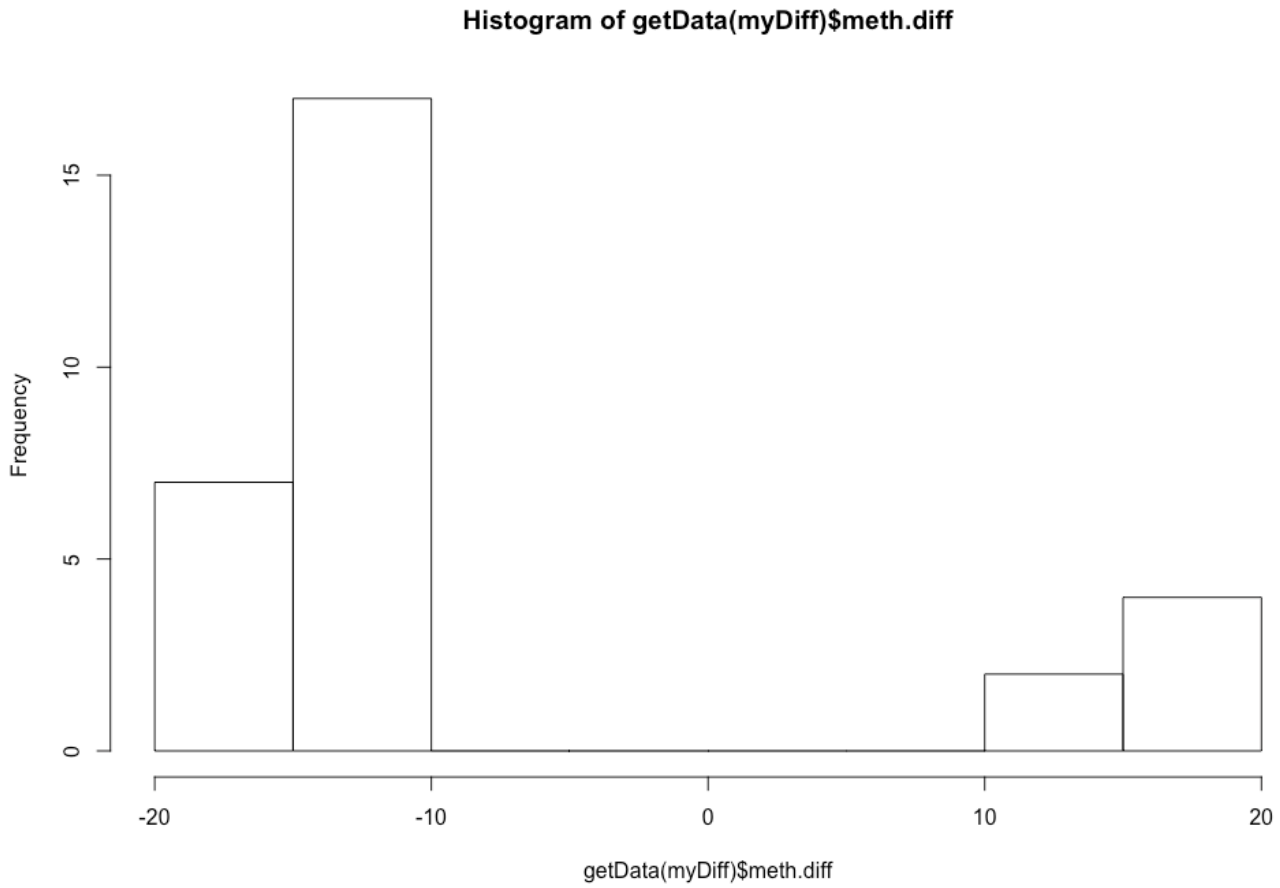
1. AA_F25 and AH_F25 (control versus high CO2)
2. AA_F25 and HA_F25 (control versus high T)

In getmethyldiff I used difference = 10

look up things about these general trends

1 table: specific genes we found
1 figure: two histograms on top of each other

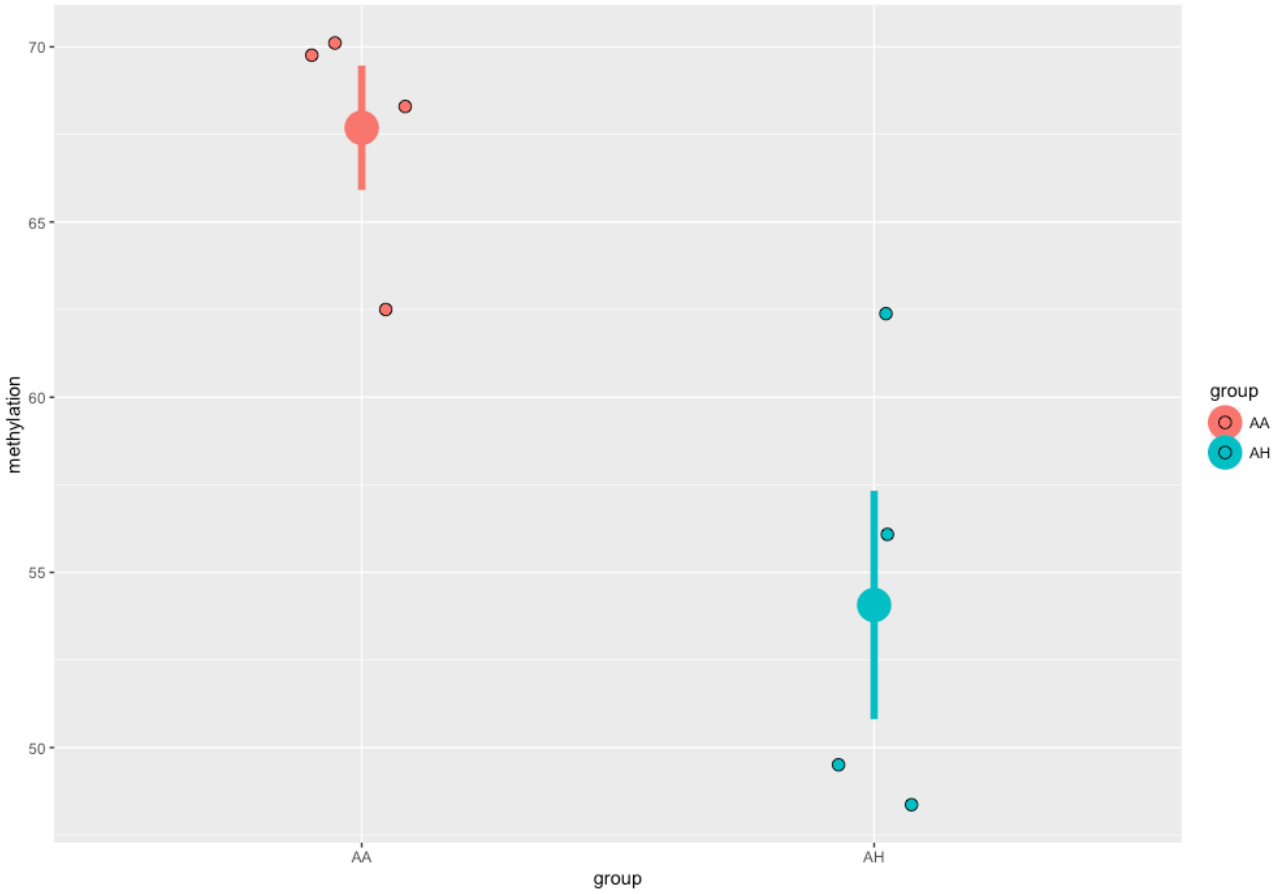
AA vs AH



30 significantly differentially methylated SNPs
most of them had reduced methylation in reponse to high CO2 by 10 to 15%

there were 6 SNPs that were hyper methylated
and 24 that were hypo methylated instead

looking at the first in that list of 30:



LS051659.1:1214

out of these 30, 18 SNPs were found to associated with biological function

unique:

15 .

1 TRINITY_DN123249_c0_g4_i4

1 TRINITY_DN139898_c0_g4_i4

1 TRINITY_DN139938_c2_g6_i1

3rd Deoxyuridine 5'-triphosphate nucleotidohydrolase, **dUTP diphosphatase activity**

This enzyme is involved in nucleotide metabolism: it produces dUMP, the immediate precursor of thymidine nucleotides and it decreases the intracellular concentration of dUTP so that uracil cannot be incorporated into DNA.

Assembled from ATP, bicarbonate and glutamine, the uracil and cytosine nucleotides are fuel for the synthesis of RNA, DNA, phospholipids, UDP sugars and glycogen.

The synthesis of UMP starts from glutamine, bicarbonate, and ATP, and requires six enzyme activities. (*De novo* pathway of uridine-5′-monophosphate (UMP) synthesis.)

various DNA damage and repair pathways and fine-tuned regulation of well-balanced deoxynucleotide (dNTP) pool work hand-in-hand
The dUTPase family of enzymes is responsible for the removal of dUTP from the nucleotide pool by hydrolyzing it into dUMP and inorganic pyrophosphate6,7,8. The importance of this enzymatic action is evident in light of the fact that most DNA polymerases cannot distinguish dUTP and dTTP and will readily incorporate the uracil-analogue if it is available in the cellular dNTP pool
dUMP as the substrate for thymidylate synthase

“was annotated as a deoxyuridine 5′-triphosphatenucleotidohydrolase which indicates a potential function in pyrimidine metabolism where HCO3−is an important factor”

higher methylation in AH gene body, so probs up-regulated in response to CO2

UMP synthesis higher because of increased bicarbonate concentration → more UTP can lead to DNA mutations where U is incorporated instead of T → needs more of this enzyme to remove UTP from the pool

heat → DNA damage → repair needs buliding blocks

2nd Endonuclease III-like protein 1, involved in base-excision repair, again seems to be related to DNA damage
lower methylation in AH! so probs downregulated...

catalyzes the first step in base excision repair (BER), the primary repair pathway for the repair of oxidative DNA damage

Scientists have demonstrated that carbon dioxide (CO2) plays a role in the formation of oxidative damage in vivo. Under conditions of oxidative stress, certain types of damage (cell death, some DNA lesions, mutation frequency, etc.) affecting the

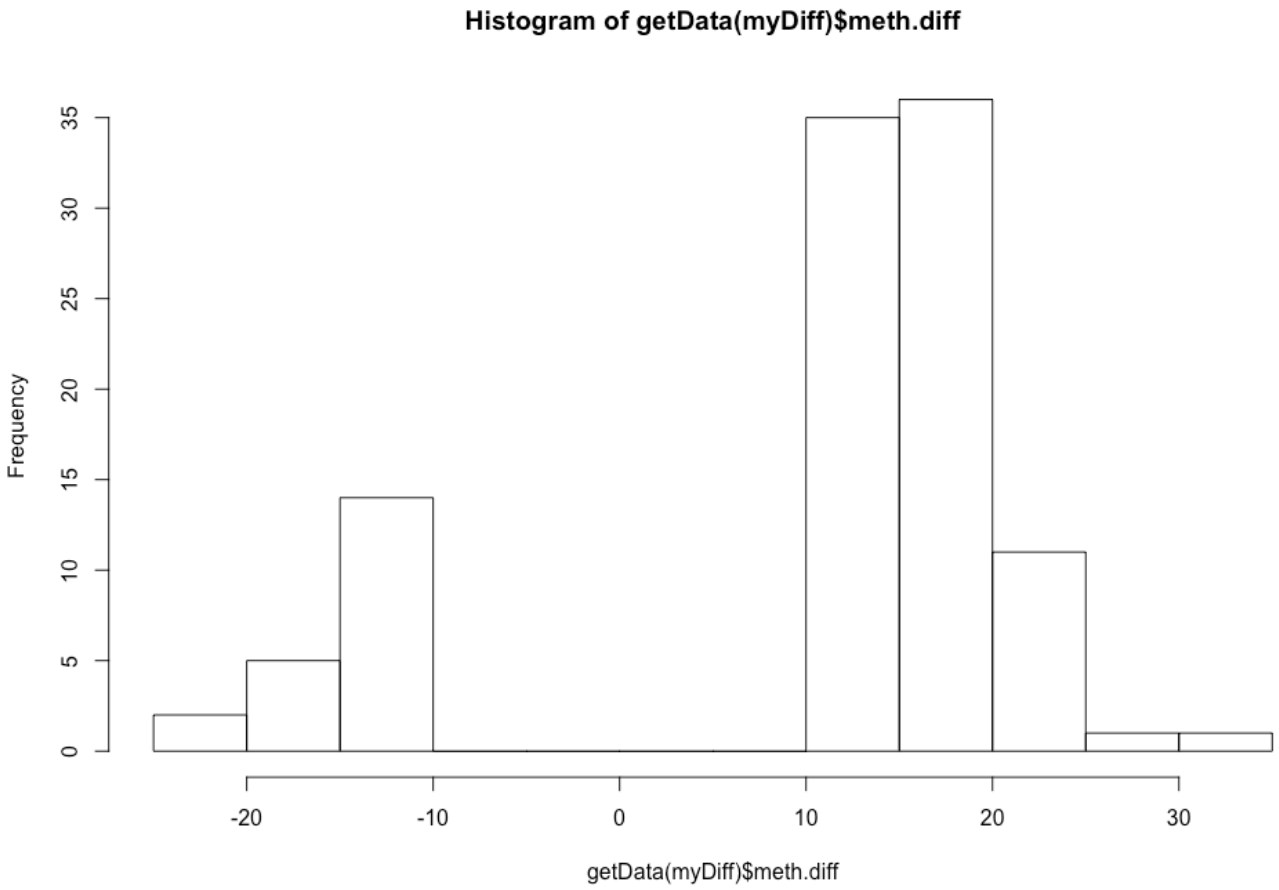
model organism Escherichia coli tend to increase depending on the level of atmospheric CO2.

but this would suggest upregulation

ncbi homologe would have this SNP in the 1st instron and that could mean upregulation
https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-018-0205-1
https://www.ncbi.nlm.nih.gov/gene/39877756
no good genomic annotation is available for this species yet
future study! where are these SNPs located! changing filtering?

1st Retrovirus-related Pol polyprotein from transposon - check methylation! i fotgot decreased methylation

AA vs HA



interestingly in high temperature SNPs are more methylated in general

hyper methylated: 84 (much more than in AH)
hypo methylated: 21 (similar to AH)
in sum: 105 (more than 3 times as much!)

76 was found with bedtools (not really, same loci could hit multiple genes)
some of these hits are in the same gene
"But I think it just means that there are multiple epigenetic changes within the same gene
so more than one loci in the same gene that has differential methylation"

if in gene body methylation → upregulation
if in promoter methylation → downregulation!!!! so check for above

Name	Methylation	Position	Function/role in heat

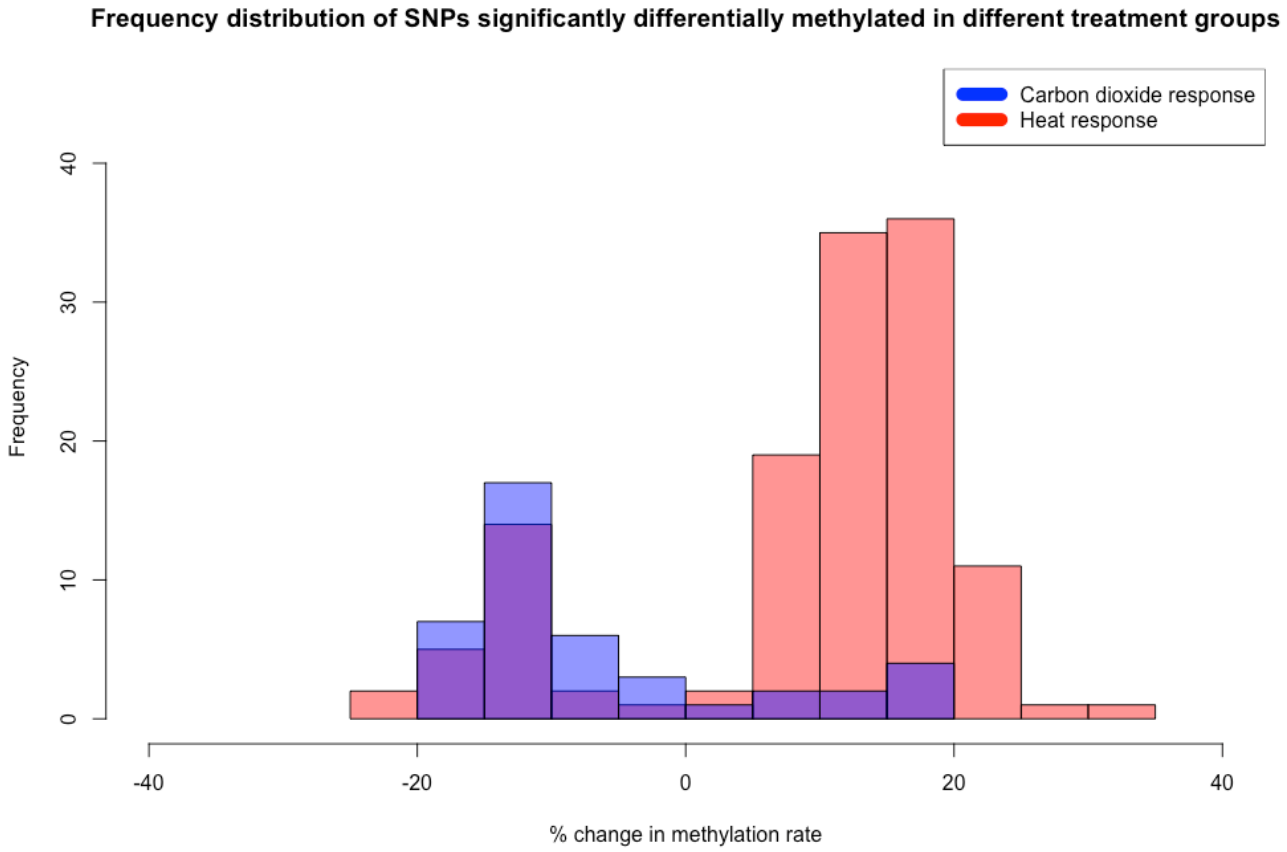
Endogenous retrovirus group K member 5 Gag polyprotein	decreased	in gene body →lower expression probably	? virus
Ral GTPase-activating protein subunit alpha-1	decreased	in gene body →lower expression probably	signal transduction acts as a *GTPase activator* for the *Ras*-like small *GTPases* RALA and RALB functions as a molecular switch to activate a number of biological processes, majorly **cell division** and transport,
Retrovirus-related Pol polyprotein from transposon opus	2 SNPs 1st increased 2nd increased	both in gene body	transposon
Transposon Ty3-G Gag-Pol polyprotein	increased	in gene body	transposon
Deoxyuridine 5'-triphosphate nucleotidohydrolase	2 SNPs 1st increased 2nd increased	both in gene body	see above
G1/S-specific cyclin-D2	increased	downstream	positive regulation of G1/S transition
Gypsy retrotransposon integrase-like protein 1	increased	downstream	transposon

2 snps not the same chr!!

on cell cycle regulation and heat induced methylation-
https://www.ncbi.nlm.nih.gov/pubmed/25712591

DNA methylation of TEs and repeats inactivates their transcription and is an evolutionary mechanism of defense against selfish DNA. Gene-body methylation was found to correlate with high expression levels

Changes in the ambient conditions may also activate transposons which are normally inactivated by epigenetic silencing through methylation of their promoters



much more in the hyper methylated category and bigger differences, absolute value max is 30.6

only 1 SNP was diff methylated in both conditions, exluding diff<10

LS068114.1	6322

which is the Deoxyuridine 5'-triphosphate nucleotidohydrolase
and it was increased in methylation both times

Conclusion (1-2 paragraphs)

- Give your biological conclusion so far from the data: What have we learned about how the environments may cause changes in methylation? What genes are involved? Explicitly use the results to address the 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.

answer questions in the begining!

very different genes are hyper/hypo methylated
more snps are diff methylated (3x) in heat response
1 overlap
other results to hilight
methylation seems to be nonrandom with respect to function → helps with response to stressful env → could give time for adaptive mutations to arise and thus fascilitate adaptation to climate change

future studies

References (listed on a separate page)

- Cite papers in MLA format.