

Subs Paper Things to Do:

- more genomes
- new outgroups? (too distant)
- explain high dS values in *B. subtilis*
- potentially poor alignment and non-orthologous genes (core genome, change methods?)
- non-parametric analysis for subs
- gap in *Escherichia coli* fig 5
- new methods for trees
- concerned about repeated genes (TEs) and not analyzing core genome
- check if trimming respects coding frame
- clear distinction between mutations and substitutions in intro (separate sections)
- datasets from previous papers (repeat my analysis on them?)
- why would uncharacterized proteins have higher subs rates?
- R^2 values in regression analysis
- update gene exp paper ref
- why are the lin reg of dN , dS and ω NS but the subs graphs are...explain!
- mol clock for my analysis?
- GC content? COG? where do these fit?

Inversions and Gene Expression Letter Things to Do:

- create latex template for paper
- confirm inversions with dot plot
- make dot plot of just gene presence and absence matrix (instead of each site) to see if this will go better
- look up inversions and small RNA's paper Marie was talking about at Committee meeting
- write outline for letter
- write Abstract
- write intro

- write methods
- compile tables (supplementary)
- write results
- write discussion
- write conclusion
- do same ancestral/phylogenetic analysis that I did in the subs paper

General Things to Do:

- summarize references 40 and 56 from Committee meeting report (Brian was asking)

Last Week

Inversions + Gene Expression:

- ✓ preliminary DESeq analysis (with toy data frame)
- ✓ Inversions and position preliminary figure
- ✓ fixed indentation issue in gene mapping code
- ✓ explored H-NS protein data

Subst Paper:

- ✓ re-submit!!! Woo!

Inversions + Gene Expression:

Queenie is still finishing things up and moving slowly. But she is still willing to do work so that is good!

As per our discussion a few weeks ago, I will be incorporating genomic position of inversions into my analysis, DESeq, and HNS proteins. I started the position and inversions analysis by looking at inverted regions that had a significant difference in normalized gene expression (not differential gene expression via DESeq, yet). There appears to be no significant correlation between distance from the origin and significant inversion blocks (blocks that had a significant difference in gene expression between inverted and non-inverted sequences using wilcox sign-ranked test). I have been trying to visualize where these significant inversions are located in the genome, but having varying genomic positions for each taxa makes this messy. My first attempt at showing significant blocks is in Figure 1. It is difficult to see what is going on because you can not tell which inversion matches the others. Please excuse the ugly axis, colours and general aesthetics. Those will be fixed. I then

attempted to just show what was happening in *E. coli* K12 MG655 (Figure 2). Figure 2 is a bit deceiving because it is actually plotting all genes within each block. My vision for this graph is sketched in Figure 3. My thought was that if I could somehow get only one genomic position associated with each inversions, then it would be easier to visualize and I could have more useful information on the graph. Since the inversions are all relative to *E. coli* K12 MG655, I thought that maybe just using its genomic positions would be ok? Or, I could find an average position for the inverted and non-inverted sequences within each block? **What are your thoughts on all this? Do you know how to best represent this data?** All positions are currently accounting for bidirectional replication and distance from the origin.

I started to run the DESeq pipeline on a toy data frame (the actual data frame is still being created by Queenie). Since we are using replicated from multiple experiments, I am accounting for these effects in the analysis. I will be running this pipeline on each of the inversion combinations (i.e. an inversion in ATCC but not any of the other taxa, an inversion in 2 out of 4 taxa...etc) and then be able to apply the results to all blocks that had that same inversion pattern.

I also realized there was a minor issue in my gene mapping code (incorrect indenting), so I am re-running this part of the analysis (the part that lists the homologous genes from the PARSNP alignment).

I looked into the H-NS protein binding data available there is quite a lot which is exciting! There are three datasets (Gainger et al. 2006, Ueda et al. 2013, and Higashi et al. 2016) that have nice CSV's with the genomic position of H-NS binding sites. There are also 2 datasets that have the same information but in cumbersome formats like PDF (Ishima et al. 2006 and Lang et al. 2007), which would be a bit more work to extract the data from. I have not read any of these papers in depth, just skimmed them. I am not sure if I should combine the information from all these datasets, or only use the intersection of the binding sites from each experiment, or only use the most recent one? the Higashi dataset also has some cutoff value to try and eliminate noise from their data set so I am not sure if I should use only that. **What are your thoughts on all of this?**

This Week

- Queenie: compare blast results and alignments
- Queenie: new dataframe for DESeq
- Keep working on position and inversion visualization
- keep looking at H-NS protein stuff: read in data and combine datasets

Next Week

- actual analysis on DESeq data
- visualizations/results for ↑

- keep working on H-NS protein stuff: combine with inversion positions
- read papers on H-NS proteins

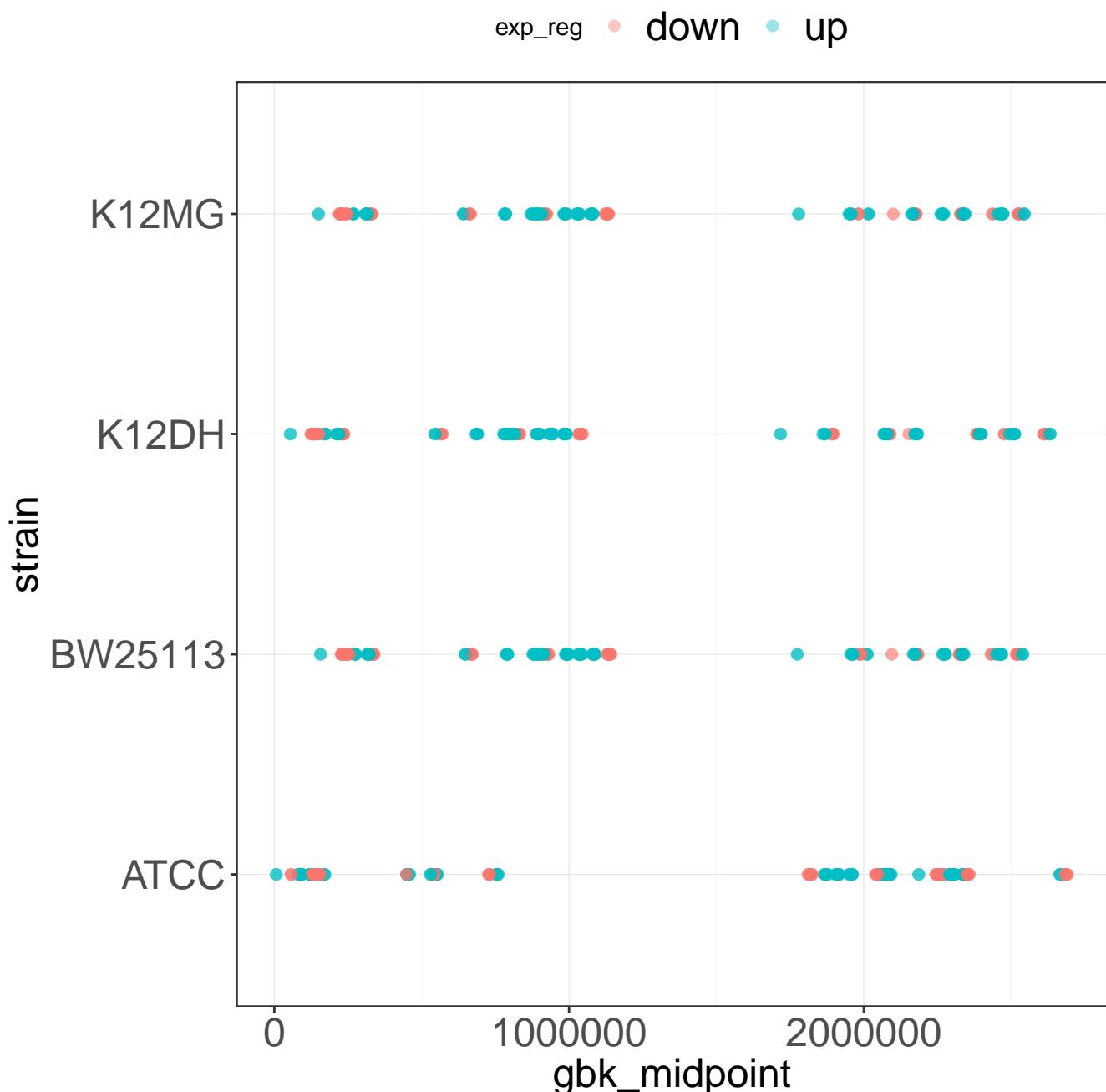


Figure 1: An attempt at showing all taxa and the significant inversions (significant = blocks that had a significant difference in gene expression between inverted and non-inverted sequences using wilcox sign-ranked test). The “up” and “down” simply mean that the sequences that were inverted in that block has significantly higher or lower gene expression than the non-inverted sequences in that block. This up/down was NOT done using DESeq, just a simple wilcox sign-ranked test.

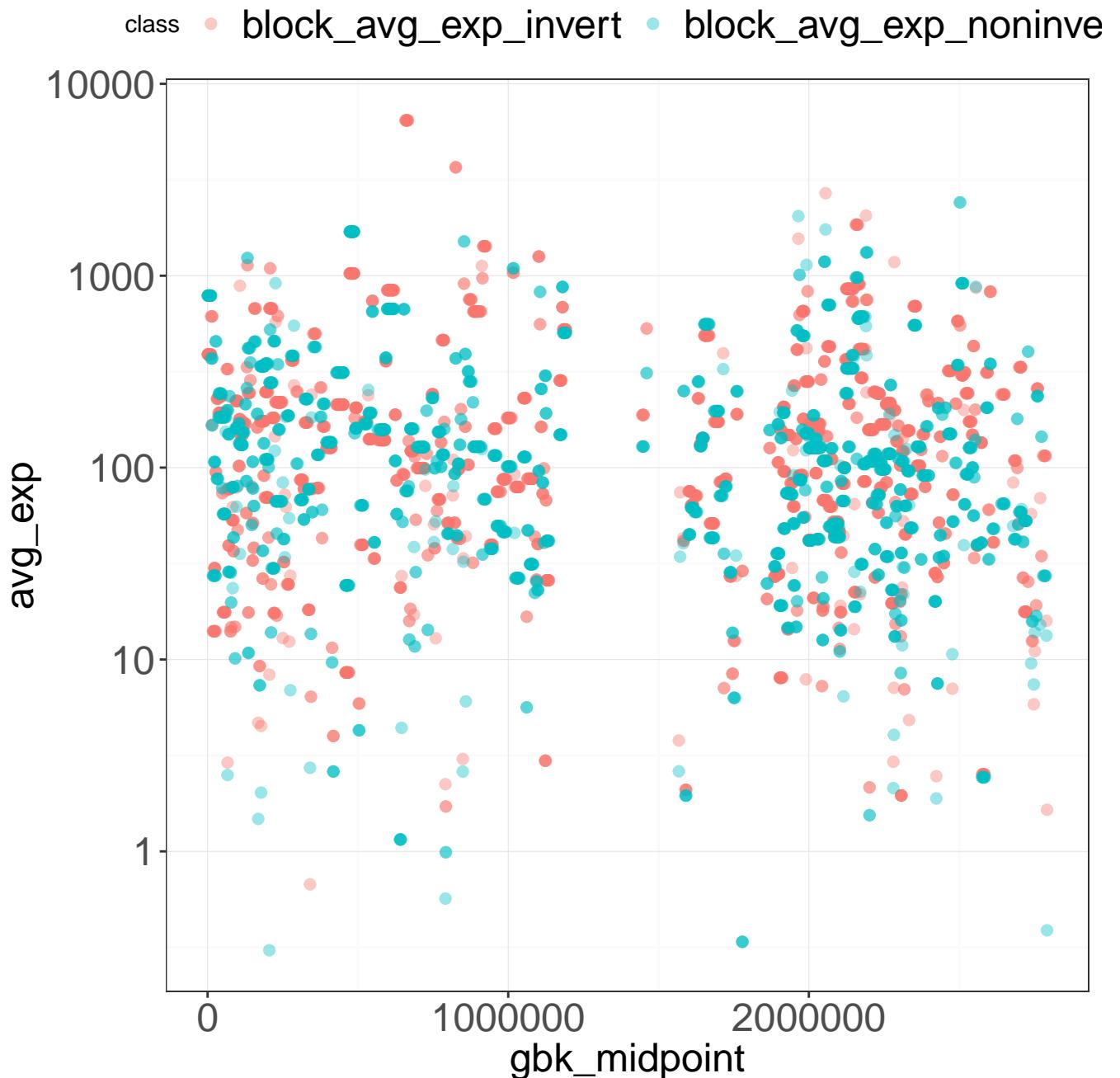


Figure 2: An attempt at showing just the significant inversions (significant = blocks that had a significant difference in gene expression between inverted and non-inverted sequences using wilcox sign-ranked test) using the distance from the origin of *E. coli* K12 MG655. The “up” and “down” simply mean that the sequences that were inverted in that block has significantly higher or lower gene expression than the non-inverted sequences in that block. This up/down was NOT done using DESeq, just a simple wilcox sign-ranked test.

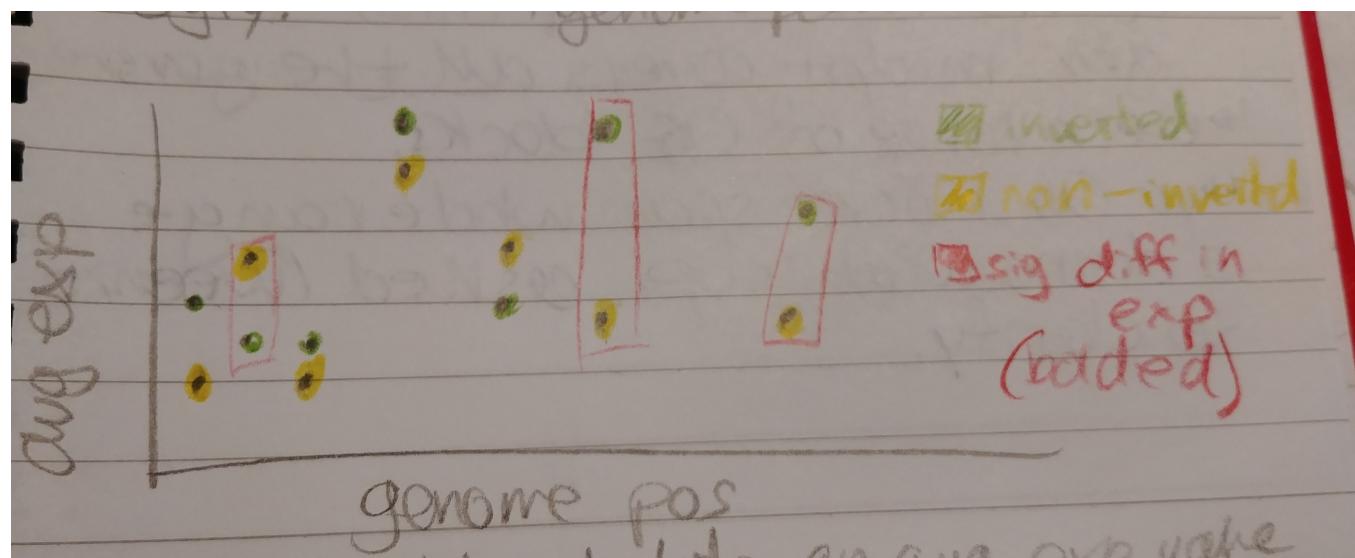


Figure 3: A sketch of what I am hoping the position graph would look like. Each inverted region/block would be plotted and the average gene expression of the inverted and non-inverted sequences would be in different colours (green and yellow). Any blocks that had a significant difference in gene expression (significant = blocks that had a significant difference in gene expression between inverted and non-inverted sequences using wilcox sign-ranked test), would be somehow bold in the diagram (red box).