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-parametirc 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:

✓Queenie: summary graphs and stats about the inversions

✓Queenie created dataframe with raw expression data and inversions info

✓Queenie created dataframes with various taxa present

✓Queenie: started to compare outputs from various blast parameters

✓preliminary analysis on dataframes with various taxa

✓started implementing origin/bidirectional scaling of the inversion/gene expression values

✓Attempt to create progressiveMauve like picture of inversions

Subst Paper:

✓chose 26 random *E. coli*, 25 random *B. subtilis*, and 26 random *Streptomyces* genomes to begin looking at

Inversions + Gene Expression:

Queenie finished creating graphs summarizing the gene expression, length and position of inverted and non-inverted genomes. However, they are not really informative and I think might just be better as tables showing the averages. It appears as though only some blocks have significant differences in those categories so to try and show all blocks at once does not visually create any stark difference between inversions and non-inversions.

Queenie also finished creating dataframes with multiple combinations of taxa present in each block so we could see if it made a difference. I have completed preliminary analysis on these dataframes and the numbers are very similar, so there might not be any huge difference in including blocks with less than all taxa (Table 1).

Queenie has started to write code to compare the reciprocal best blast hits from the various parameters I chose. There were some interesting issues she brought up to me such as not all blast gene names are present in the genbank file. I am not sure what is going on here but the gene names just do not exist in the genbank file for some genes. **any thoughts on this? Should I be overly concerned?**

I started to implement the same origin/bidirectional scaling code used in my other papers and noticed that the strand information was missing from the dataset. I am getting Queenie to look into this.

I tried to make a progressiveMauve-like picture of the genomic inversions but had trouble. The PARSNP output can not be read into progressiveMauve (still looking into this but it does not look possible) so I can not use progressiveMauve to create the diagram. The visualization produced by PARSNP is not as useful because it does not clearly connect blocks that have been shuffled around the genome. I tried to look into a program called GenomeDiagram which was used in one of the Galardini papers on *Sinorhizobium*. However, this required an updated version of BioPython on the machines and only reads in genbank files (from my knowladge) so I am not sure how to convert the PARSNP outputs to create the "features" needed to create the same diagram. The other option (which would be easiest) is to just use R to plot the x-y coordinates from matching genes (either from the blast output or my gene mapping code). **What are your thoughts on all of this?**

Substitution Paper

I think that the reason that the analysis looks "fine" for the 23 *S. meliloti* genomes is that they are quite similar, and therefore aligning them does not pose any issues. So I decided to test out how long the alignment would take and how much would be trimmed with 25 random (complete) genomes from *E. coli*, *B. subtilis*, and *Streptomyces*. I have updated the progressiveMauve time plots with the new values for *B. subtilis* and *E. coli* (Figures 1 and 2). The results are still roughly the same, that having about 60 sequences would take over a month to align. The alignments do appear to be "messy" with 96 block for *E. coli* (Figure 3) and 113 blocks for *B. subtilis* (Figure 4).

I do not have a plot of how long the individual block re-alignments with `MAFFT` take (**I can make one if you think it is useful?**) but it takes about 5h for the longest *E. coli* block to align.

Overall, I think that this provides some evidence that the more diverged the taxa are, the more messy the analysis becomes.

This Week

- Queenie: compare blast results and gene mapping
- Queenie: combine gene mapping and gene info (from gbk)

- continue scaling up the *Escherichia coli*, *B. subtilis*, and *Streptomyces* genomes
- think about (and execute) how to incorporate distance from the origin into the inversion analysis
- work on DESeq pipeline
- find new outgroups for subst trees
- assess Queenie's results from various blast parameters

Next Week

- Queenie to create a plot of the inversions
- Queenie: determine which RBBH's are the same as the homologous genes identified from the alignments
- keep working on increased number of genomes for subst analysis
- distinction between mutations and substitutions in subst paper intro

	% of Blocks that are		
Datasets: Taxa Per Block	Inverted	Inverted and Differentially Expressed	Increased in Gene Expression in the Inverted Sequences
All 4	68.15	8.66	60.61
At least 3	68.23	8.91	57.14
At least 2	68.02	8.96	58.33

Table 1: Percent of blocks in categories for various datasets (blocks with all 4 taxa, at least 3 taxa, or at least 2 taxa). The second column is any block that had at least one sequences that was inverted. The last column only deals with blocks that had at least one inverted sequence and had a significant difference in gene expression (column 3).

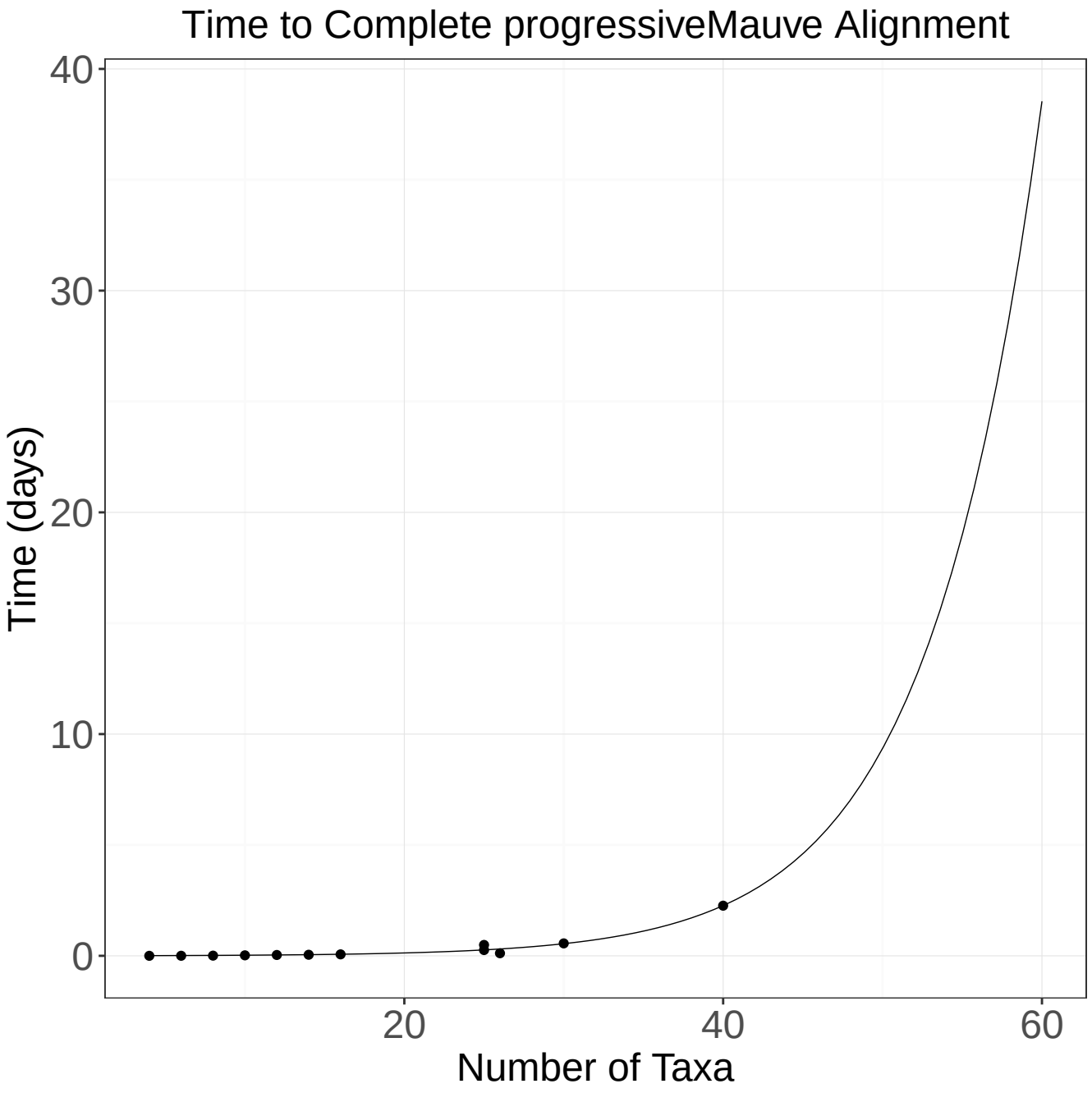
Time to Complete progressiveMauve Alignment
Time (days)
0
10
20
30
40
20
40
60
Number of Taxa

Figure 1

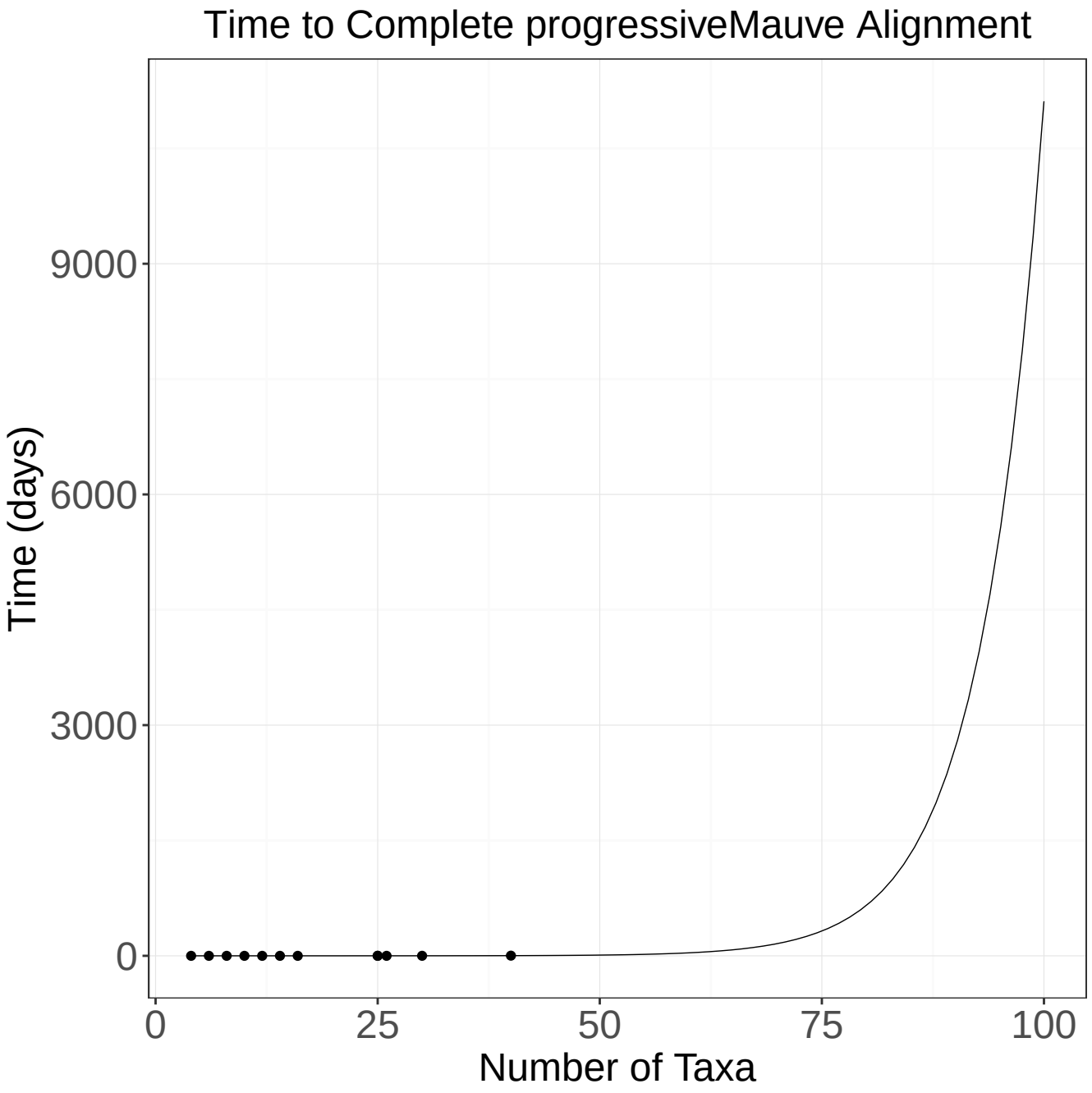
Time to Complete progressiveMauve Alignment
Time (days)
9000
6000
3000
0
0
25
50
75
100
Number of Taxa

Figure 2


Escherichia coli str. K-12 substr. MG1655
Escherichia coli BL21(DE3)
Escherichia coli BW25113 K-12
Escherichia coli ECC-1470
Escherichia coli O182:H21 D181
Escherichia coli 108
Escherichia coli 112
Escherichia coli CI5
Escherichia coli EcPF16
Escherichia coli KBN10P04869


Bacillus subtilis subsp. subtilis str. 168
Synechocystis sp. PCC 6803
Bacillus subtilis BSn5
Bacillus subtilis ATCC 21228
Bacillus subtilis SP1
Bacillus subtilis subsp. inaquosorum DE111
Bacillus phage SPbeta

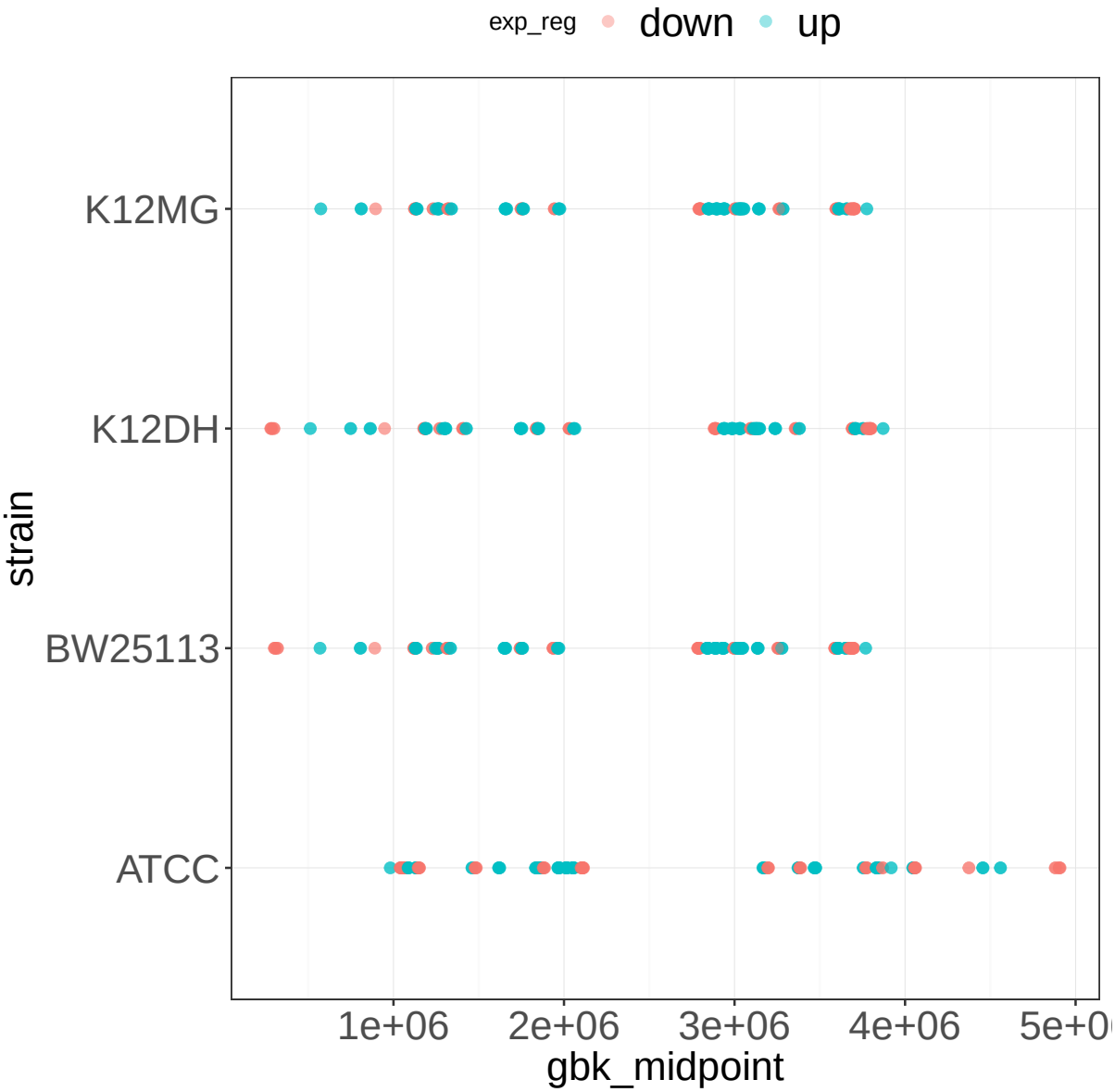

Figure 5: Test graphic for the visualization of inversions and distance from the origin of replication. Each dot represents a gene in a block where there is a significant difference in gene expression between inverted and non-inverted sequences within that block. The points are coloured based on if the inverted sequences have higher expression ("up") or lower expression ("down") compared to the non-inverted sequences. Genomic position is on the x-axis with NO bidirectional replication accounted for.