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: comparing blast and gene alignment homologs

✓Queenie: start creating dataframe that is compatible with `limma`

Subst Paper:

✓re-running subst code (matching codon positions) for 25 *E. coli* and *B. subtilis* genomes

✓re-running subst analysis (6 genomes per bac) with new trees

✓started non-parametric analysis (subs + dN, dS, and ω)

✓previous datasets from other papers

✓new *Streptomyces* analysis with 5 genomes (2 new *S. coelicolor* genomes)

Inversions + Gene Expression: Queenie has begun comparing the blast output and the alignment homologs, however there are lots of issues with missing gene names/IDs and not being able to have a reliable comparison. We are still working through this.

Substitution Paper

I realized that with more genomes, there are more nuances with my substitutions code (the one that ensures we are comparing the same codon position). I am currently working out these kinks. However, based on what I am seeing so far, it looks like a lot of the sites are trimmed. So the more genomes you have, the worse progressiveMauve is at catching homologous blocks, creating an unreliable alignment. I will be working on coming up with quantitative numbers for this theory.

I am re-doing the subst analysis with the new phylogenetic trees (from RAxML). The subst part of the analysis is at the regression stage, and the dN, dS, and ω portion is still at the gene alignment phase.

The new *Streptomyces* analysis (with the two new *S. coelicolor* genomes) is under way and at the same stage as the other bacteria above.

One reviewer suggested that I use the same datasets from previous papers (Sharp et al., 2005, Cooper et al., 2010, Flynn et al., 2010, Morrow and Cooper 2012) to see if I am still seeing the same trends. I made a table of those studies below outlining which taxa they use and how many genomes (Table 1). Most of these papers are dealing with completely different taxa than I am, and most of the genomes are complete reference genomes. I am not sure if it is worth it or necessary for me to do my analysis on these datasets to appease this reviewer. **What are your thoughts?**

This Week

- Queenie: compare blast results and alignments
- Queenie: new dataframe for `limma`
- write about TEs and repeated elements in cover letter
- re-run inversions mapping based on correct subst code
- re-run subst analysis for 25genome bac based on correct subst code
- non-parametric analysis for subs analysis
- quantify trimming loss (more subst genomes)
- selection analysis for subst analysis (6 genomes per bac)

Next Week

- Queenie: new dataframe for `limma`
- previous subst papers datasets (can I re-do?)
- why do uncharacterized proteins have higher sub rates?
- gap in *E. coli* fig 5
- *B. subtilis* high dS values should not be present
- blast to confirm homologs in subst analysis
- distinction between mutations and substitutions in subst paper intro

Authors	**Species Name**	**Genomic Structure**
Galardini et al. 2013	*Sinorhizobium* (14 genomes)	1 chromosome, 2 plasmids
Morrow et al. 2012	*Burkholderia* (4 genomes)	3 chromosomes
Cooper et al. 2010	*Burkholderia* (6 genomes) *Bordetella* (4 genomes) *Vibrio* (4 genomes) *Xanthomonas* (5 genomes)	3 chromosomes single circular chromosome 2 chromosomes single circular chromosome
Couturier et al. 2006	126 species of bacteria (only chromosome)	only primary chromosome for multi-repliconic bacteria
Flynn et al. 2010	*Sulfolobus* (6 genomes)	single circular chromosome
Sharp et al. 2005	80 species of bacteria (genomic)	only primary chromosome for multi-repliconic bacteria

Table 1: Previous studies looking at substitutions/mutation rate.

	% 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 2: 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).

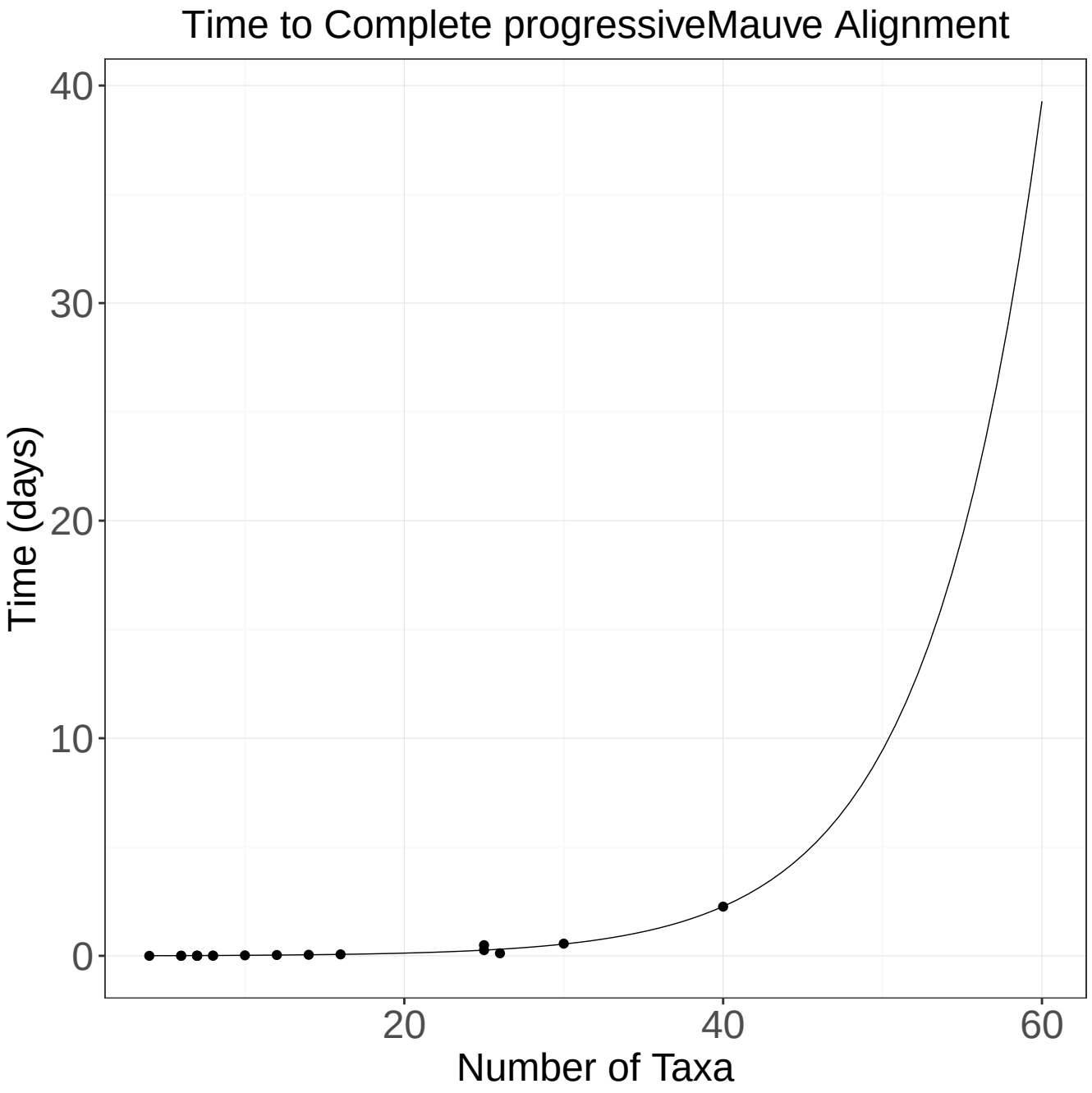

Time to Complete progressiveMauve Alignment
Time (days)
0
10
20
30
40
20
40
60
Number of Taxa

Figure 1

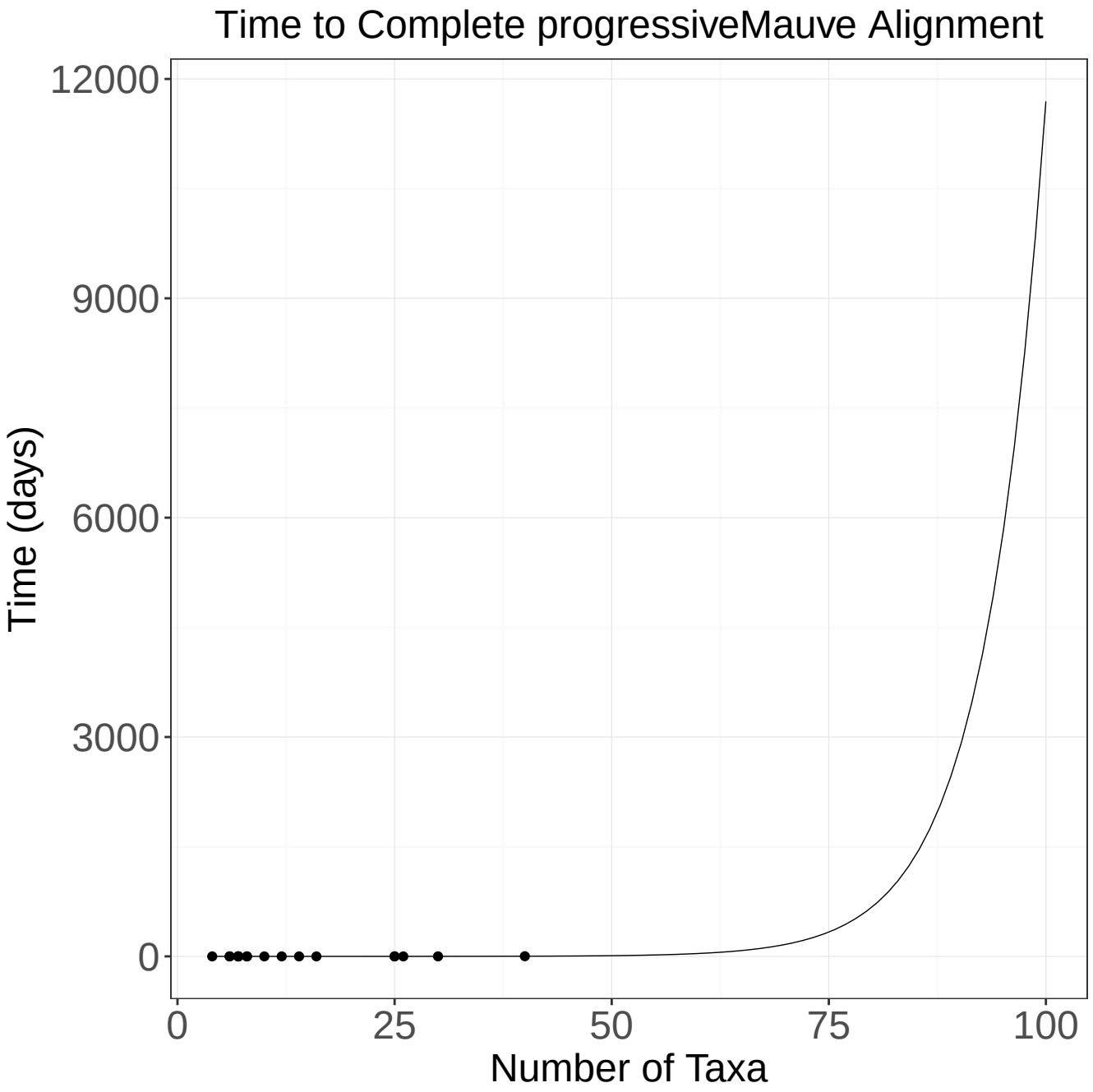
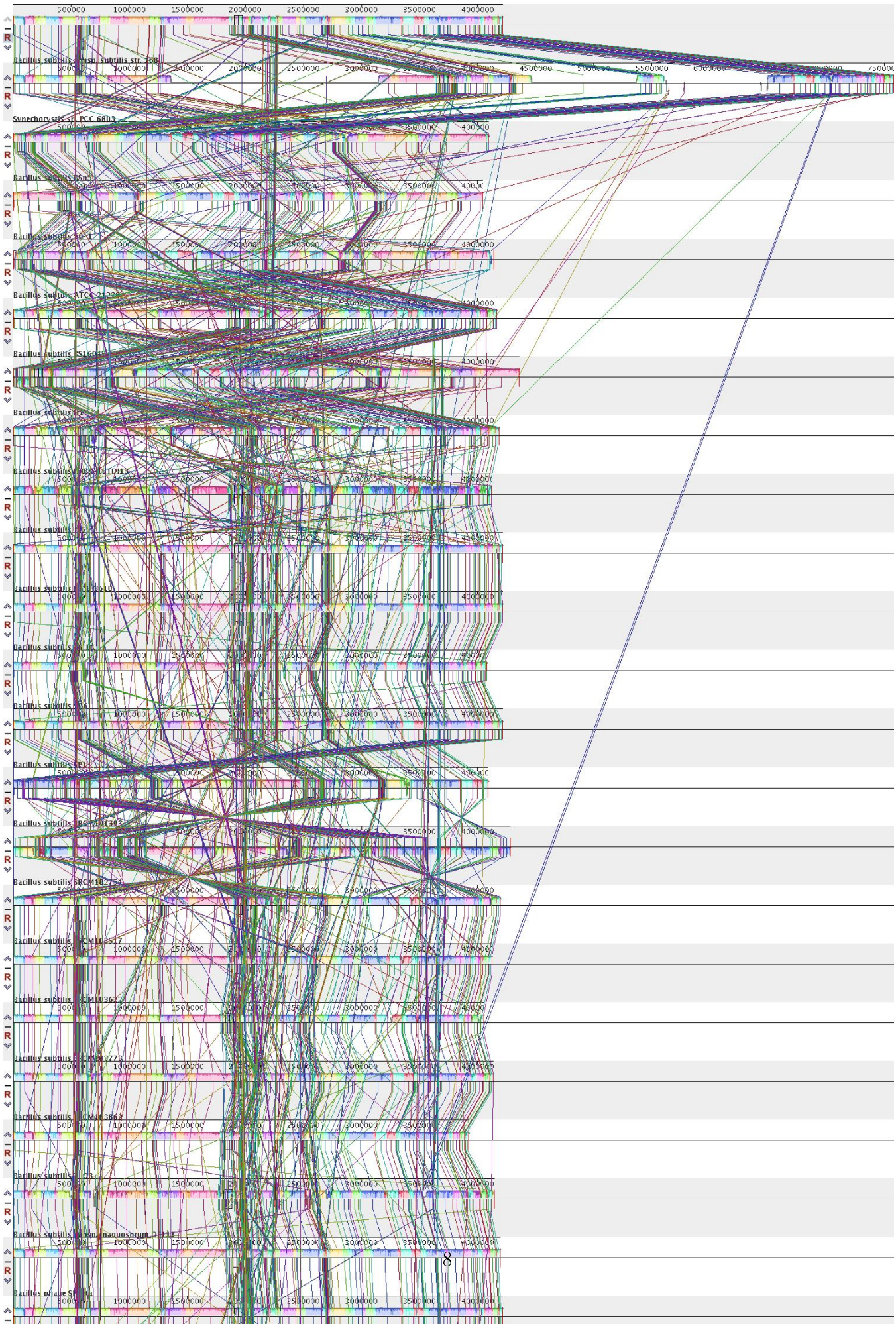

Time to Complete progressiveMauve Alignment
12000
9000
6000
3000
0
Time (days)
0
25
50
75
100
Number of Taxa

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 4/1-1
Escherichia coli 4/4
Escherichia coli CI5
Escherichia coli Ecol_224
Escherichia coli EcPF16
Escherichia coli O2:K1 ExPEC XM
Escherichia coli KBN10P04869
Escherichia coli RHBSTW-00
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
5000000
5500000


Bacillus subtilis subsp. subtilis str. 168
Synechocystis sp. PCC 6803
Bacillus subtilis BSn5
Bacillus subtilis QB-1
Bacillus subtilis ATCC 21332
Bacillus subtilis BS168
Bacillus subtilis
Bacillus subtilis QB928 HTDH13
Bacillus subtilis
Bacillus subtilis
Bacillus subtilis
Bacillus subtilis
Bacillus subtilis SP1
Bacillus subtilis
Bacillus subtilis BRCM1027
Bacillus subtilis BCMT8817
Bacillus subtilis BCM10362
Bacillus subtilis BCM03773
Bacillus subtilis BCM13862
Bacillus subtilis
Bacillus subtilis subsp. natto BEST195
Bacillus phage SPO1

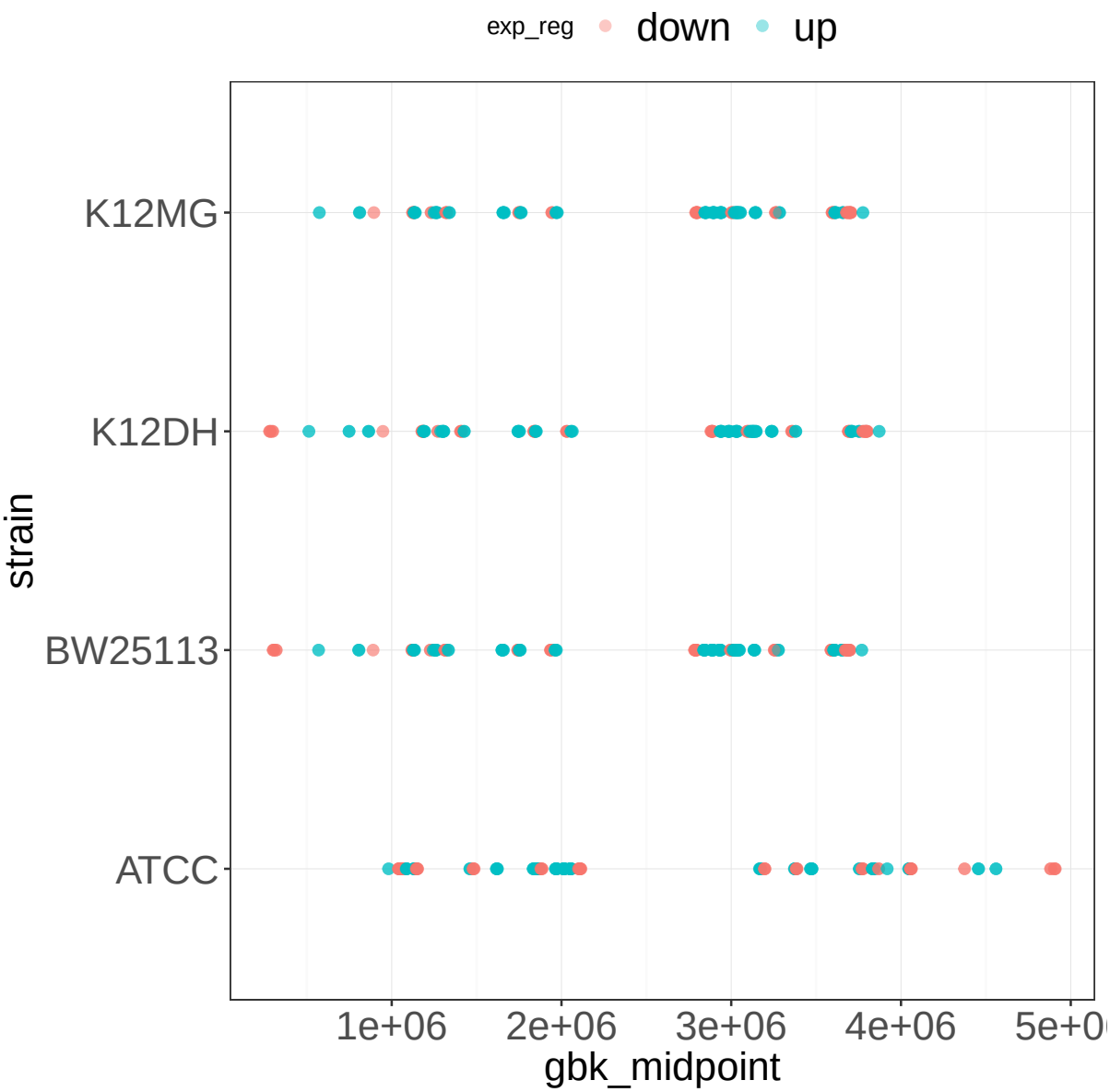


Figure 5: Test graphic for the visualization of inversions and distance from the origin of replication. Each dot represents a gene in a block where there is a significant difference in gene expression between inverted and non-inverted sequences within that block. The points are coloured based on if the 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.