Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of ↑~~
- ~~Look into *Streptomyces* non-coding issue~~
- ~~Look into *E. coli* coding issue~~
- ~~Look into pSymB coding/non-coding trend weirdness~~
- ~~Figure out why *Streptomyces* appears to have tons of coding data missing~~
- ~~Figure out what is going on with cod/non-cod code and why it is still not working!~~
- ~~write up methods for coding/non-coding~~
- ~~write methods and results for clustering~~
- ~~start code to split alignment into multiple alignments of each gene~~
- figure out how to deal with overlapping genes
- figure out how to deal with gaps in gene of ref taxa
- split up the alignment into multiple alignments of each gene
- get dN/dS for coding/non-coding stuff per gene
- Or get 1st, 2nd, 3rd codon pos log regs
- write up coding/non-coding results
- take out gene expression from this paper
- write better intro/methods for distribution of subs graphs
- mol clock for my analysis?
- write discussion for coding/non-coding
- GC content? COG? where do these fit?
- write coding/non-coding into conclusion

Gene Expression Paper Things to Do:

- ~~look for more GEO expression data for *S. meliloti*~~
- ~~look for more GEO expression data for *Streptomyces*~~
- ~~look for more GEO expression data for *B. subtilis*~~

- ~~format paper and put in stuff that is already written~~
- ~~look for more GEO expression data for *E. coli*~~
- ~~Get numbers for how many different strains and multiples of each strain I have for gene expression~~
- ~~re-do gene expression analysis for *B. subtilis*~~
- ~~re-do gene expression analysis for *E. coli*~~
- ~~find papers about what has been done with gene expression~~
- ~~read papers ↑~~
- put notes from ↑ papers into word doc
- write abstract
- write intro
- add stuff from outline to Data section
- create graphs for expression distribution (no sub data)
- add # of genes to expression graphs (top)
- average gene expression
- write discussion
- write conclusion
- add into methods: filters for Hiseq, RT PCR and growth phases for data collection
- update supplementary figures/file

Inversions and Gene Expression Letter Things to Do:

- ~~get as much GEO data as possible~~
- ~~find papers about inversions and expression~~
- ~~see how many inversions I can identify in these strains of *Escherichia coli* with gene expression data~~
- ~~read papers about inversions~~
- check if opposite strand in progressiveMauvemeans an inversions (check visual matches the xmfa)
- check if PARSNP and progressiveMauveboth identify the same inversions (check xmfa file)
- create latex template for paper

- put notes from papers into doc
- use large PARSNP alignment to identify inversions
- confirm inversions with dot plot
- 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

## Last Week/Holiday

dN/dS Super confused. This book I found that basically walks you through how to get dn/ds says that you need to first find orghologs and then align them because codeml requires codon based alignments (duh bc this is how it can actually distinguish between syn and non-syn subs and why I was getting weird errors when I tried to just put my aln into paml). So what I am wondering is should I be starting from scratch and finding orthologs with blast and then going from there? Or should I still use the blocks specified by mauve but re-align them so they are codon alignments? Do I use gene trees or my same whole genome tree? (maybe whole genome tree is fine bc I already showed that the trees of each block is very similar to the trees of the overall genome and blocks that wernt were removed) Do I vary the dn/ds over sites? branches? both?

✓read papers about gene expression trends in bacteria

✓read papers about inversions in bacteria

✓apply for a few more scholarships

I spent the holiday break reading papers and making notes. These papers were about trends in gene expression in bacteria (more general) and inversions in bacteria (general) as well as the impact on inversions and rearrangements on gene expression in bacteria.

For the papers solely on gene expression, they were all basically saying the things that we already know, that gene expression tends to decrease when moving away from the origin. However, there were varying explanations as to why this was happening. Some papers attributed this to an increase in gene dosage near the origin, but others said that this was not the case and there must be some other mechanism for controlling this or that gene expression trends could be a secondary effect of selection on say gene order or chromosome organization. But, they all basically said that because bacterial genomes are so highly organized based on gene order, physical folding of the chromosome, co-regulation of gene clusters...etc that this is why we see gene expression decreasing when moving away from the origin of replication.

For the papers on Inversions, the results were a bit all over the place. Most of the papers were older (1980's) and often focused on one known inversion in one bacteria. These often were related to antibiotic resistance, flagella state of bacteria, or turning specific genes on/off. Sometimes these studies would say how the inversion altered expression of close down or upstream genes by changing promoters locations. Other studies engineered their own inversions, some mentioning gene expression and some don't and talk more about the replication of inversions or how they tend to be symmetrical around the origin. One large study done with Staphylococcus in 2012 just mentions that lots of genes are deferentially expressed (some up regulated and some down regulated). The overall feeling I got was that inversions can alter gene expression and this can be done by changing the promoter location of specific genes, which can impact many things about the bacteria like their growth state and resistance. There seems to be no "trend" with respect to inversions always causing gene expression to go up or down. It also looks like no one has done what we are doing: looking at existing expression data to see how having an inversion impacts gene expression within and outside of the inversion.

I also applied for a few more scholarships over the break. I will keep you updated if I get any of them!

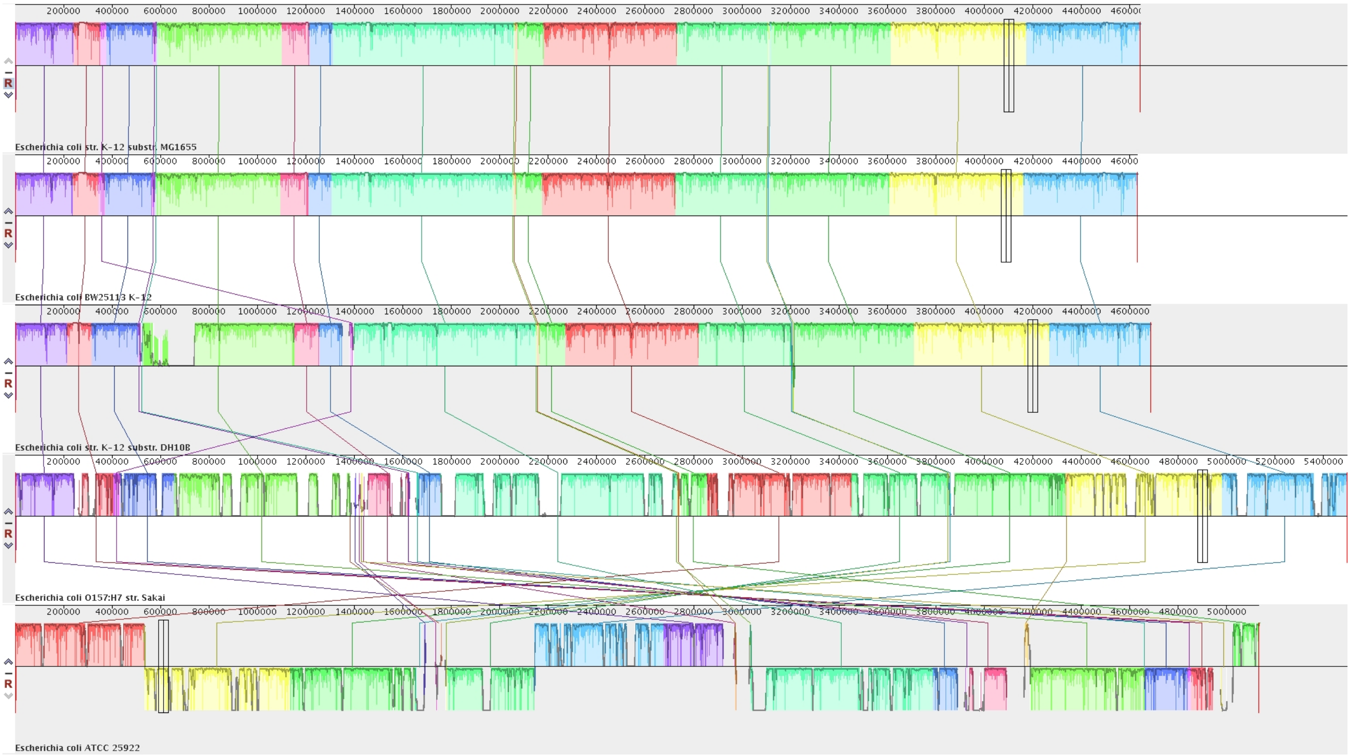
I have also been working on the code that will split up each of the blocks into an alignment of each gene so that I can calculate dN/dS for each of these genes.

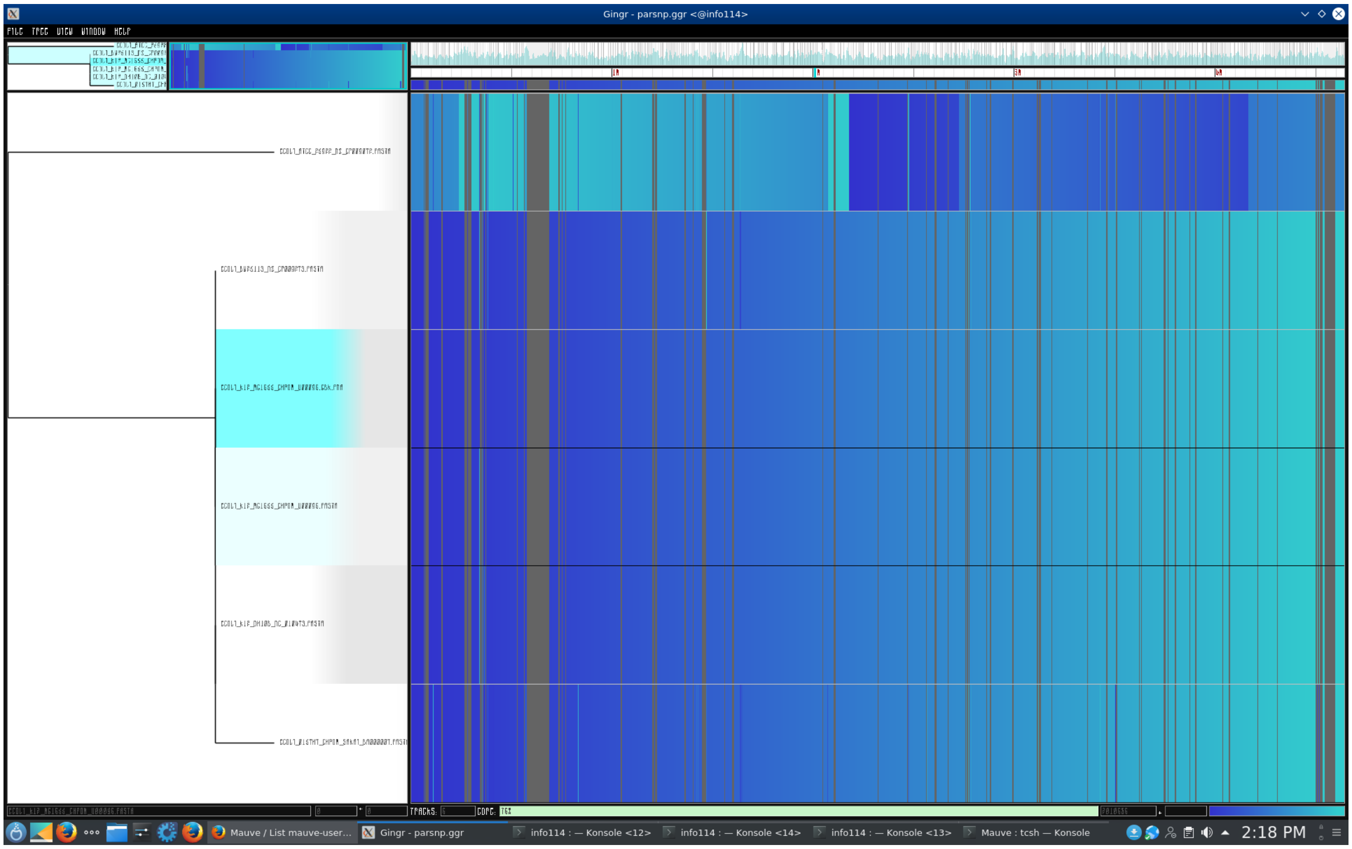
## This Week

I would like to finish making the code for printing out the alignment of each gene so that I can then calculate the dN/dS for each gene of all the bacteria. I still need to work out how to deal with overlapping genes, how to deal with gaps in the reference sequence, and how to print out the alignment of each gene in a readable format I would like to read 2 more papers this week.

## Next Week

I would like to have parameters for the PAML dN/dS calculation figured out so I can then run this on each gene for each bacteria. I want to begin figuring out how to obtain all inversions from the Mauve or PARSNP alignment.


Escherichia coli str. K-12 substr. MG1655
Escherichia coli BW25113 K-12
Escherichia coli str. K-12 substr. DH10B
Escherichia coli O157:H7 str. Sakai
Escherichia coli ATCC 25922
200000
400000
600000
800000
1000000
1200000
1400000
1600000
1800000
2000000
2200000
2400000
2600000
2800000
3000000
3200000
3400000
3600000
3800000
4000000
4200000
4400000
4600000
4800000
5000000
5200000
5400000
R


Gingr - parsnp.ggr <@info114>
File
Tree
View
Window
Help
ECOLI_ATCC_25922_NZ_CP009072.FASTA
ECOLI_BW25113_NZ_CP009273.FASTA
ECOLI_K12_MG1655_CHROM_U00096.GBK.FNA
ECOLI_K12_MG1655_CHROM_U00096.FASTA
ECOLI_K12_DH10B_NC_010473.FASTA
ECOLI_O157H7_CHROM_SAKAI_BA000007.FASTA
TRACKS:
CORE:
Mauve / List mauve-user...
Gingr - parsnp.ggr
info114 : — Konsole <12>
info114 : — Konsole <14>
info114 : — Konsole <13>
Mauve : tcsh — Konsole
2:18 PM

| Bacteria and Replicon | % of Coding Sequences | % of Non-Coding Sequences | % of Subs Coding | % of Subs Non-Coding |
|---|---|---|---|---|
| *E. coli* Chromosome | 86.47% | 13.53% | 5.00% | 8.96% |
| *B. subtilis* Chromosome | 87.49% | 12.51% | 7.31% | 6.42% |
| *Streptomyces* Chromosome | 89.03% | 10.97% | 13.74% | 14.91% |
| *S. meliloti* Chromosome | 86.27% | 13.73% | 0.19% | 0.22% |
| *S. meliloti* pSymA | 83.34% | 16.66% | 2.84% | 4.58% |
| *S. meliloti* pSymB | 88.81% | 11.19% | 2.78% | 3.44% |

Table 1: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

| Bacteria and Replicon | Coding Sequences | Non-Coding Sequences |
|---|---|---|
| *E. coli* Chromosome | $-5.938\times10^{-8}$*** | $-9.237\times10^{-8}$*** |
| *B. subtilis* Chromosome | $-7.584\times10^{-8}$*** | NS |
| *Streptomyces* Chromosome | $5.483\times10^{-7}$*** | $9.182\times10^{-9}$*** |
| *S. meliloti* Chromosome | $-1.448\times10^{-6}$*** | $-7.037\times10^{-7}$*** |
| *S. meliloti* pSymA | $-9.704\times10^{-7}$*** | $-1.464\times10^{-7}$*** |
| *S. meliloti* pSymB | $5.007\times10^{-7}$*** | NS |

Table 2: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectionality of replication. All results are marked with significance codes as followed: $< 0.001$ = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', $> 0.05$ = 'NS'.

| Bacteria Strain/Species | GEO Accession Number | Date Accessed |
|---|---|---|
| *E. coli* K12 MG1655 | GSE60522 | December 20, 2017 |
| *E. coli* K12 MG1655 | GSE73673 | December 19, 2017 |
| *E. coli* K12 MG1655 | GSE85914 | December 19, 2017 |
| *E. coli* K12 MG1655 | GSE40313 | November 21, 2018 |
| *E. coli* K12 MG1655 | GSE114917 | November 22, 2018 |
| *E. coli* K12 MG1655 | GSE54199 | November 26, 2018 |
| *E. coli* K12 DH10B | GSE98890 | December 19, 2017 |
| *E. coli* BW25113 | GSE73673 | December 19, 2017 |
| *E. coli* BW25113 | GSE85914 | December 19, 2017 |
| *E. coli* O157:H7 | GSE46120 | August 28, 2018 |
| *E. coli* ATCC 25922 | GSE94978 | November 23, 2018 |
| *B. subtilis* 168 | GSE104816 | December 14, 2017 |
| *B. subtilis* 168 | GSE67058 | December 16, 2017 |
| *B. subtilis* 168 | GSE93894 | December 15, 2017 |
| *B. subtilis* 168 | GSE80786 | November 16, 2018 |
| *S. coelicolor* A3 | GSE57268 | March 16, 2018 |
| *S. natalensis* HW-2 | GSE112559 | November 15, 2018 |
| *S. meliloti* 1021 Chromosome | GSE69880 | December 12, 2017 |
| *S. meliloti* 2011 pSymA | NC_020527 (Dr. Finan) | April 4, 2018 |
| *S. meliloti* 1021 pSymA | GSE69880 | November 15, 18 |
| *S. meliloti* 2011 pSymB | NC_020560 (Dr. Finan) | April 4, 2018 |
| *S. meliloti* 1021 pSymB | GSE69880 | November 15, 18 |

Table 3: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.