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~~
- ~~check if each gene alignment is a multiple of 3 (proper codon alignment)~~
- 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~~
- ~~write discussion for coding/non-coding~~
- ~~write coding/non-coding into conclusion~~
- make a list of what should be in supplementary files for subs paper
- put everything in list into supplementary file for subs paper
- write dN/dS methods
- write dN/dS results
- write dN/dS discussion
- write dN/dS into conclusion

- mol clock for my analysis?
- GC content? COG? where do these fit?

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 progressiveMauve means an inversions (check visual matches the xmfa)
- check if PARSNP and progressiveMauve both 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

✓figure out how to deal with overlapping genes ✓figure out how to deal with gaps in gene of ref taxa ✓check if each gene alignment is a multiple of 3 (proper codon alignment) ✓split up the alignment into multiple alignments of each gene ✓write better intro/methods for distribution of subs graphs ✓take out gene expression from subs paper ✓write up coding/non-coding results ✓write discussion for coding/non-coding ✓write coding/non-coding into conclusion

I spent last week writing up the coding and non-coding stuff into my substitutions paper including the methods, results, and discussion. I also mostly worked on my code for splitting up the alignment into multiple alignments for each gene. This is totally done!! It deals with gaps in the reference sequence (which messes up the genome position counting), gaps in general, and when genes are overlapping. I also got it to print the alignment in a nice pretty format that PAML can

then use! I briefly started working on actually running CODEML and using it to determine the dN and dS rates for each gene. I was getting errors with my file format so it must be something silly I am missing.

Below is a summary of what has been thrown out because of reasons (for one particular Block with an alignment length of 252,304 which includes gaps):

Category	Number of Genes	Proportion of Genes
Total Genes	188	100%
Genes not divisible by 3	50	27%
Genes with all gaps	34	18%
Total Usable Genes	104	55%

Questions for you about dN/dS:

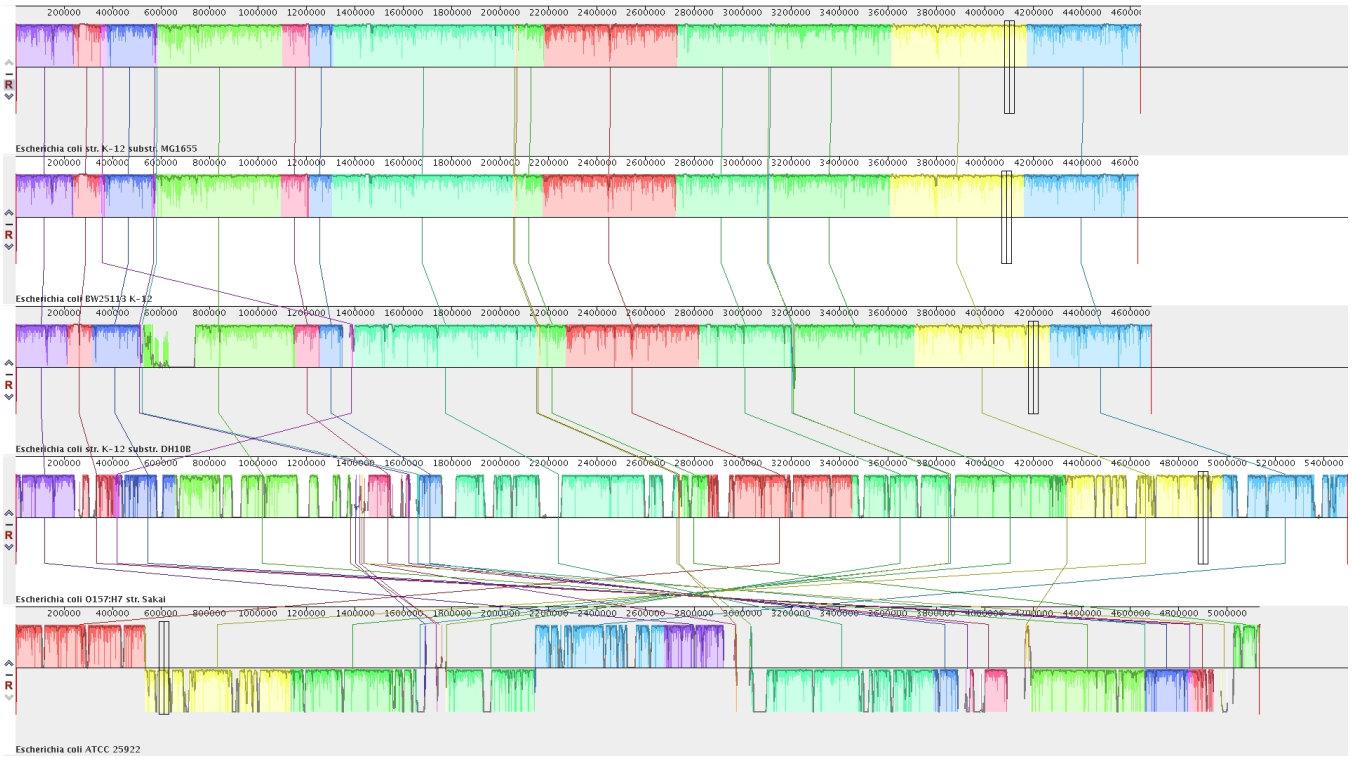
1. Should I just throw out any genes that are not divisible by 3?
2. Should I use a gene tree for each gene or the overall species tree(the one I use for the rest of my analysis)? I think the species tree is fine?
3. I have really confused myself about gaps and I really don't know if removing them is good... so I need to talk to you about this
4. Should I be allowing substitutions to vary per site? branch? or use a site-branch model?
5. I think I should be trying the simplest model and then comparing it to more complicated models and do a likelihood ratio test to see which one fits better? (I think that this can be done in PAML or it easily gives you the info and then I can do the actual test in R) If I should do this...do I do this on a per gene basis? What if a different model fits all the genes? Is this comparable?

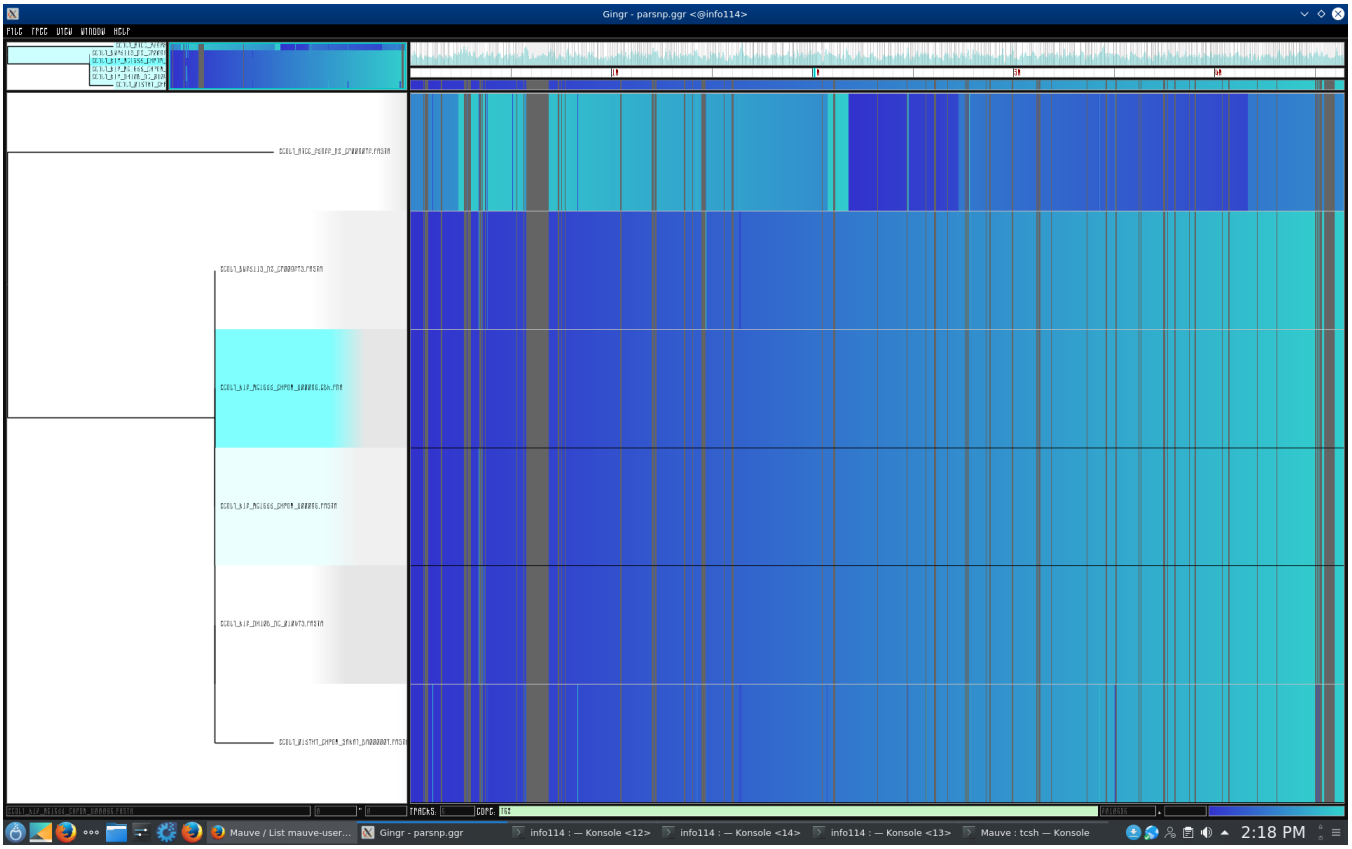
This Week

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


Gingr - parsnp.ggr <@info114>
FILE TREE VIEW WINDOW HELP
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	-5.938×10^{-8}***	-9.237×10^{-8}***
B. subtilis Chromosome	-7.584×10^{-8}***	NS
Streptomyces Chromosome	5.483×10^{-7}***	9.182×10^{-9}***
S. meliloti Chromosome	-1.448×10^{-6}***	-7.037×10^{-7}***
S. meliloti pSymA	-9.704×10^{-7}***	-1.464×10^{-7}***
S. meliloti pSymB	5.007×10^{-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.