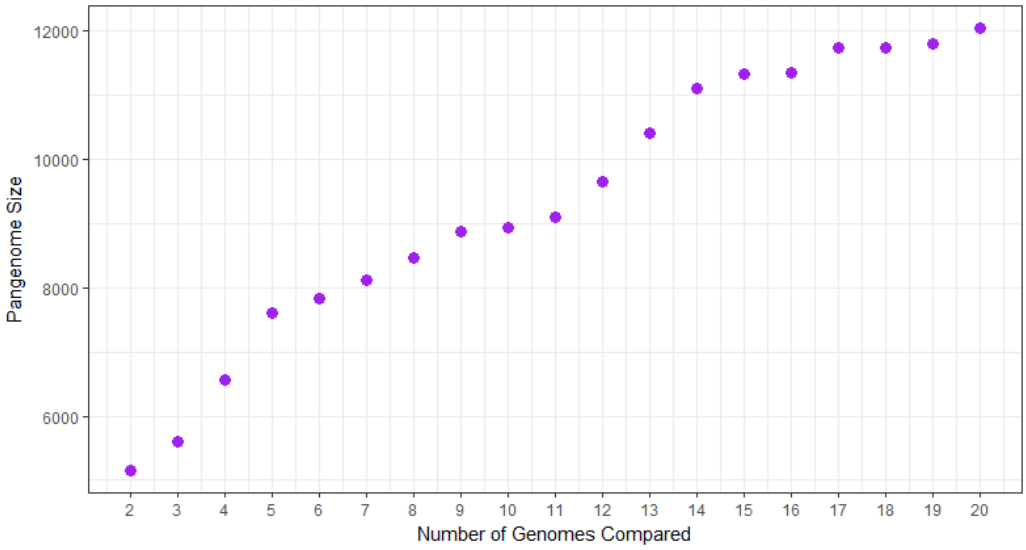
Last Week

- Mostly error proofing, accounting for variations in genome annotation and how it effects the results
- Dilemma with how to deal with *S. aureus* and how to deal with plasmids
- Verified linear regressions of *E. coli* and *Salmonella*, yet to figure out what to actually do with them and whether the linear regression is the best test to perform on the phylogenetic depth vs. pangenome size data
- Both *E. coli* and *Salmonella* proved low positive correlation between tree depth (distance measured by summing branch lengths) and pangenome size
- Completed sample error test for *E. coli* in Roary, where a steady increase was shown but more sample points (20 for each?) have to be completed, in reference to:



This Week

S. aureus Genome Annotation

First, I wanted to check what the effect was with leaving and removing plasmids from the bacterial files, and whether the annotation does indeed change the results by a large factor. Of course, when plasmids were left in, the generated trees were useless and consisted of very few branches (which makes sense). So instead I opted for sorting out the files which only had the genome sequence in them, and comparing them with my own modified files, where I simply used a code to cut out all the plasmid sequences.

The results were as followed:

	Total Number of Genomes	Pangenome Size Range
Raw Genome Only Files	63	4550 - 5705
Modified Genome Only Files	91	3774 - 4620

This was a surprising result to me, as I expected the Modified files to have a larger pangenome size overall, simply due to the larger number of genomes to choose from. Note that these modified files would've included the raw files. The only theory right now is that perhaps the addition of the modified files allowed for more genes to the core, though the fact that these ranges barely overlap has to be taken note of and looked at again in the future. For now, I will be looking using the modified data for most bacteria.

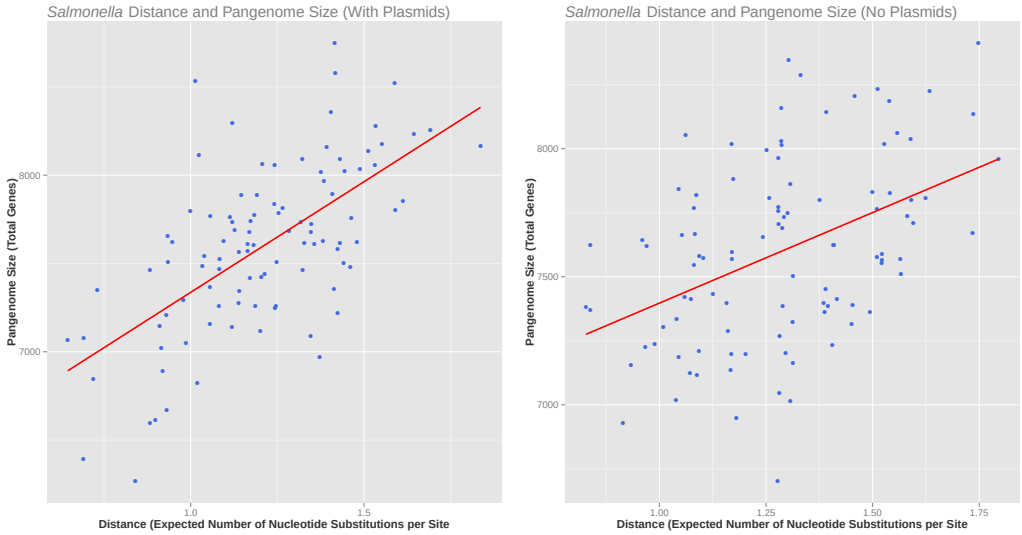
Note that the Total Number of Genomes that could be used were decided Parsnp alone, as indicated by the taxa they kept in after running the program on the full directory with all relevant files within. So these results highly depend on what Parsnp decides to "throw away". That being said, there were only 64 genomes from the downloaded files that were annotated without plasmids, hence parsnp decided to throw away only 1, whereas modified files had an original count of 153.

Salmonella Data With Plasmids Vs. Without Plasmids

This is just to see if removing the plasmids has any significance difference on the data result, so that it can be acknowledged in the future.

The following displays Distance's Correlation with Pangenome Size for raw, unsorted *Salmonella* data (meaning there was likely plasmid data mixed into it) as compared to data that was modified with plasmids all removed through a Python code (so there could technically be errors):

(Please zoom in, I still struggle with these graphics but I'm working on it I promise)



Again, this also doesn't really make sense intuitively. One would expect the data with plasmid to correlate to higher distance or larger pangenome size values since plasmids should(?) mean more variation. Yet the data with lower plasmid seems to have a lower correlation overall. This could again be rationalized by the way parsnp chooses with genomes to remain in the tree: the larger variance of data may have been completely thrown away, leaving for a more concise correlation between the remaining samples. Again this is just a theory, and I'll probably have to look deeper into how parsnp works at this rate.

Originally, 262 unmodified *Salmonella* genome files were given to Parsnp, leaving a tree with 171 tips.

For the modified files (plasmid sequences removed), 262 modified files were again given to Parsnp but leaving only 98 tips.

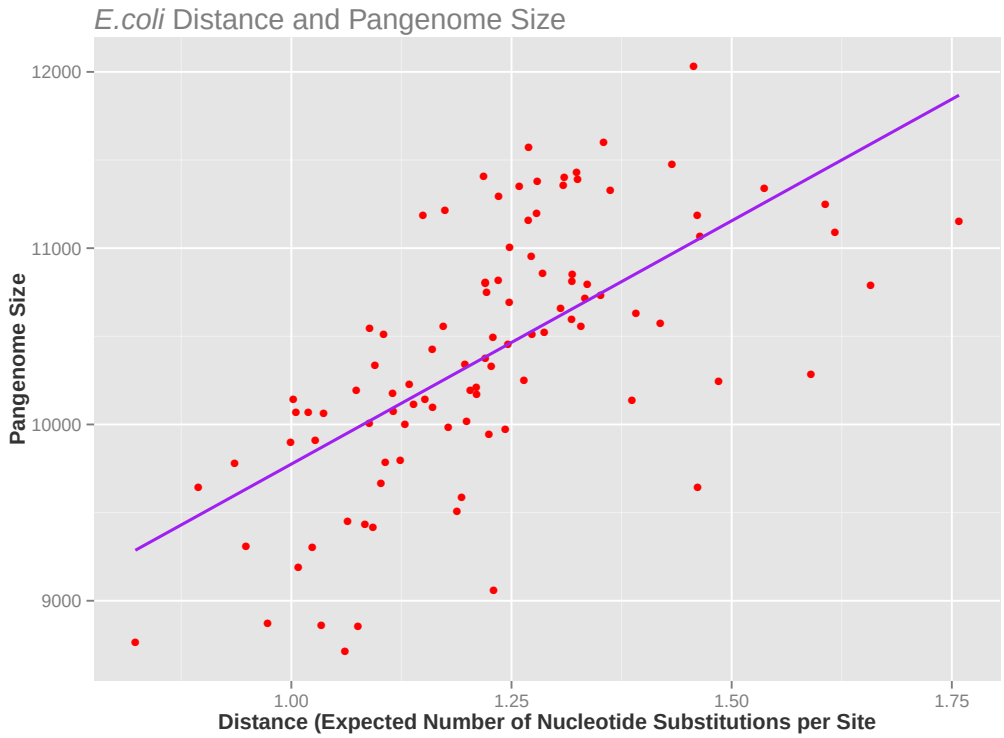
Modification of Raw Genome Files to Remove Plasmid Sequences

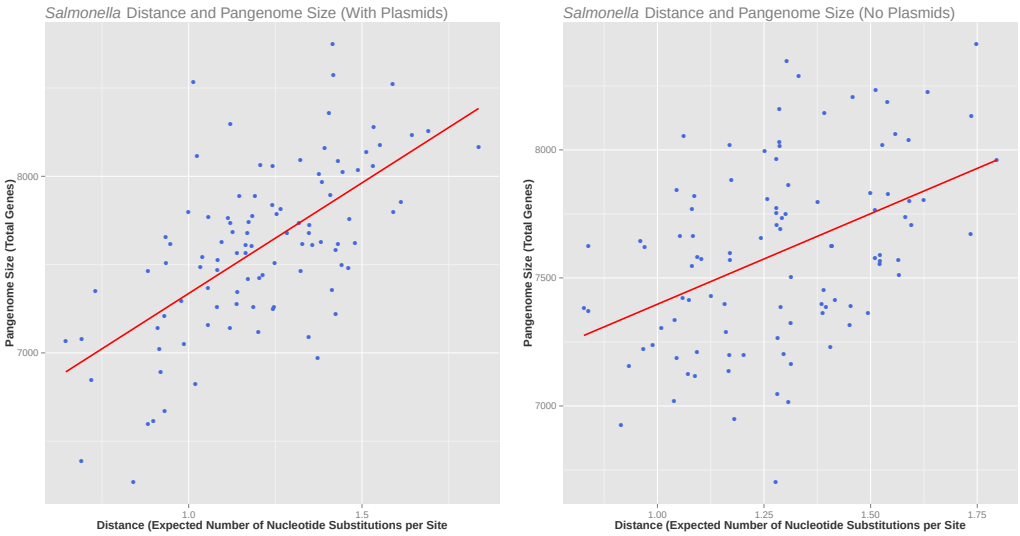
Since this may the underlying cause of all the strange results I am receiving, I will report my method of modifying these files. For most of the genomes, I verified that the "complete genome" data was recorded first, followed by the plasmid sequences. For the few files that were exception to these, I modified them manually and added them to the bunch afterwards. Knowing that each sequencing indicated its start with a ">" symbol, I used to the following Python code to simply extract the first chunk of course which correlated to the first read only:

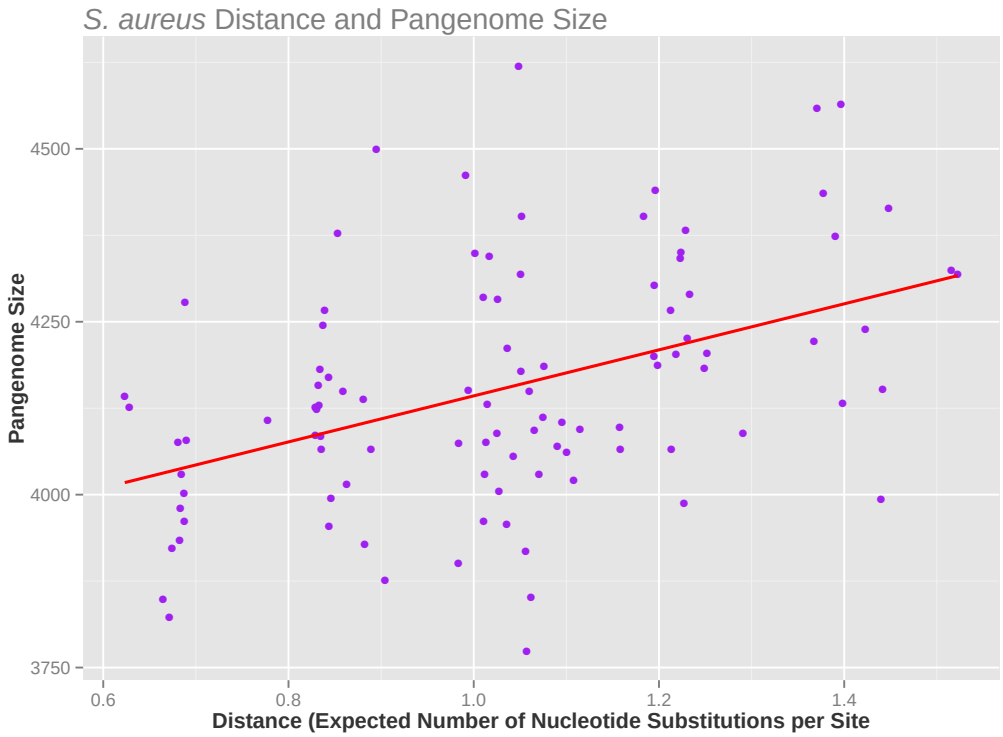
```
import sys
file = open(sys.argv[1],'r')

read = ""
new_read = ""
title_location = []
i = 1

for line in file:
 read = read + line
 for character in read:
  if character == "\n":
   i+=1
  if character == ">":
   title_location.append(i)

if len(title_location) > 1:
 a = 1
 for x in read:
  new_read = new_read + x
  if x == "\n":
   a+=1
  if a >= title_location[1]:
   break
 print(new_read)
else:
 print(read)
```

Cumulative Pangenome vs Depth Data

The following graphs display the data collected so far regarding genome distance and pangenome size for the bacterial species: *E. coli*, *Salmonella* and *S. aureus*:



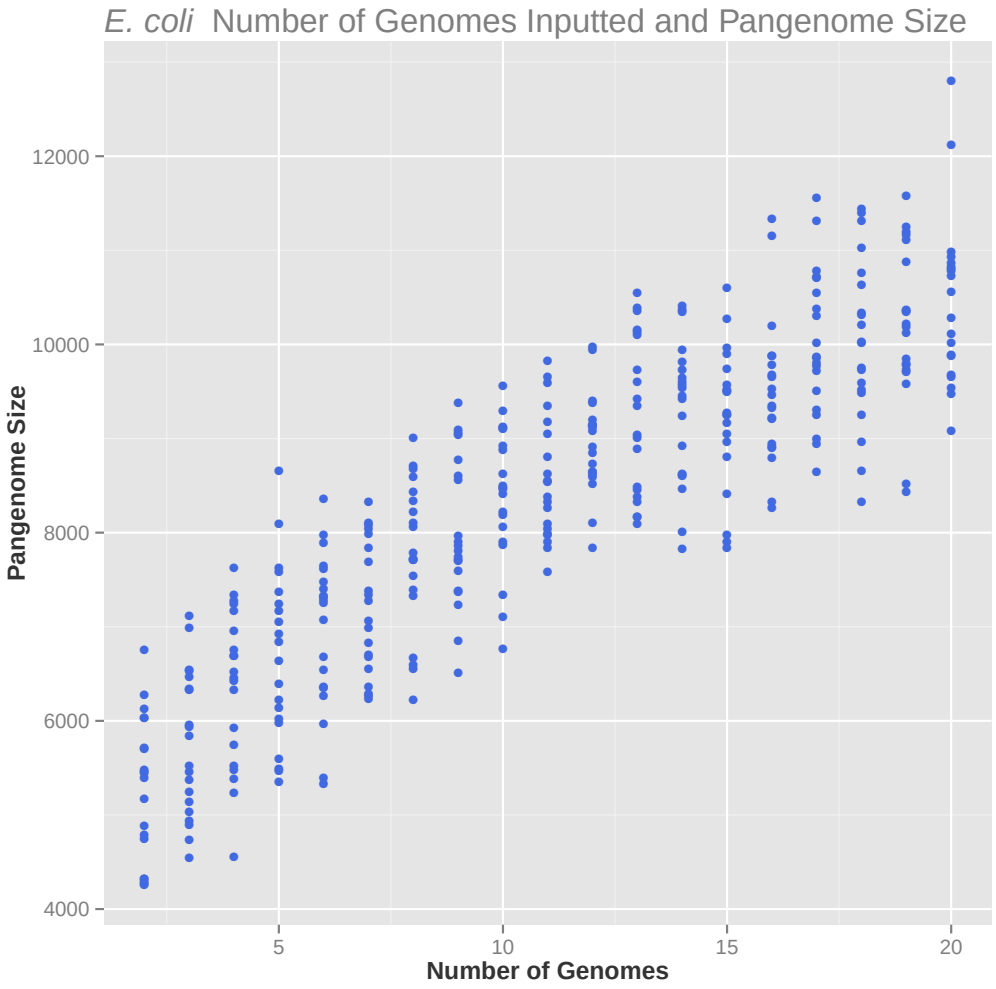




Note both the *E. coli* and *S. aureus* data have no plasmids in them.

"Rarefaction" Curve Data:

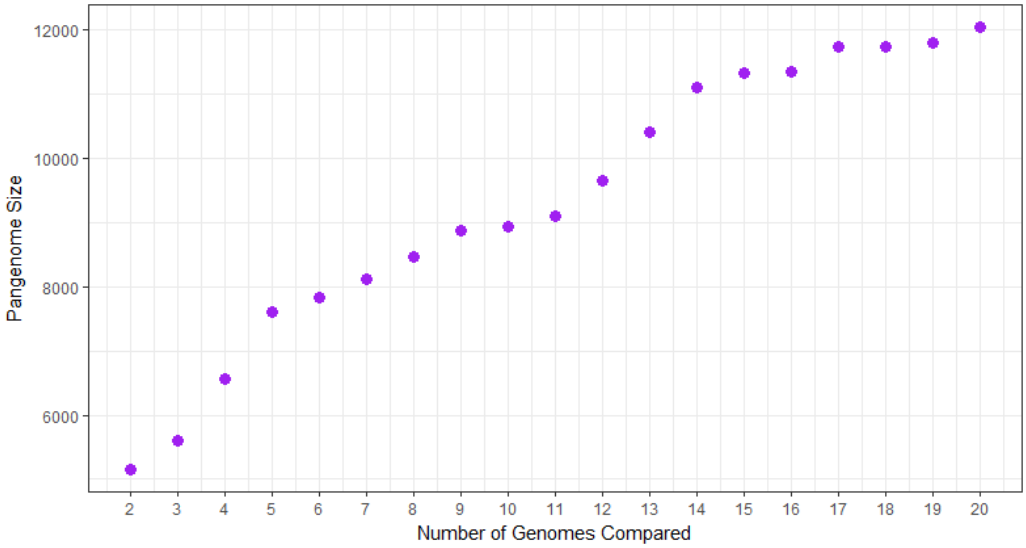
Not sure if I did this the way you wanted it, but this is the data for *E. coli* that was observed after runs with 20 combinations of 2,3...20 input genomes randomly selected from the total samples each time. All 380 runs were "independent" from each other, with its own unique individual batch corresponding to the number of input genomes required.



We see the upward sloping trend we expect, but I realize that this may not be the ideal way of doing is as I approached it. Before, the single error

curve was done "cumulatively": two genomes were randomly selected, then a third was added to the previous two and so on, finally stopping at the 20 which included all of the previous samples selected. This time, everything was random, as mentioned above.

However, there's a strange correlation in that the funny looking "dip" that occurs seems to be around the 10 genome area just as with the singular run from before:



Next Week

- Begin looking at lateral gene transfer for *E. coli*
- Look over all these codes and redo them several times because they're...iffy
- Complete Genome Number and Pangenome Size correlation data for *Salmonella* and *S. aureus*
- Start completing tests on more bacteria