Inversions and Gene Expression Paper Revisions:

Permutations: Gene expression

I did permutations tests shuffling gene expression and inversion status to see if the means differ between inverted and non-inverted regions. I did this for all blocks, looking at the overall difference between inverted and non-inverted regions. The result was not significant (which is opposite from what the wilcoxon test said). I also did a permutation test on a per-strain basis and found no difference between inverted and non-inverted regions and genes for each strain (which is opposite from what the wilcoxon test said for ATCC). This to me means that generally there is no difference between inverted and non-inverted regions. But, on a gene level, we do see difference between some inverted genes (see below, and Wilcoxon tests per block). **Overall, do you think this is still worth stating? Or since we only have a small number of inverted genes that have a significant difference in expression (8% of inverted genes), should I change the tone of the paper to say that inversions only seem to have some gene specific effects?**

However, when I did a permutation test with just the non-inverted regions of ATCC and all other strains, the test was not significant. Indicating that the non-inverted regions of ATCC have no expression difference than non-inverted regions of other strains. Which means that any difference we do see is not just due to ATCC but the inversions! Yay!

I did not do a permutation test comparing inverted and non-inverted genes within each block (this would be hundreds of tests). **Do you think that I should do a permutation test per block? Is this too much? Can we just stick with the Wilcoxon test results per block?**

Ancestral Inversion

I ran PARSNP on a few different close outgroups: *E.fergusonii*, *E. coli* Saki, *E. coli* K5198, *E. coli* TW. The other strains are *Escherichia coli* K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922. **Do you think is is “close” enough as an outgroup choice for inversions?**

I ran this analysis and here are the results:

E. fergusonii

- 17.7% of blocks had outgroup = K-12 MG = ATCC
- 31.8% of blocks had outgroup = K-12 MG
- 39.4% of blocks had outgroup = ATCC
- 11% of blocks had the outgroup with a different sign than both ATCC and K-12 MG

***E. coli* Saki**

- 36.2% of blocks had outgroup = K-12 MG = ATCC
- 56.4% of blocks had outgroup = K-12 MG
- 5.1% of blocks had outgroup = ATCC
- 2.1% of blocks had the outgroup with a different sign than both ATCC and K-12 MG

***E. coli* K5198**

- 32.4% of blocks had outgroup = K-12 MG = ATCC
- 31.3% of blocks had outgroup = K-12 MG
- 32.3% of blocks had outgroup = ATCC
- 3.9% of blocks had the outgroup with a different sign than both ATCC and K-12 MG

***E. coli* TW**

- 4.3% of blocks had outgroup = K-12 MG = ATCC
- 7.8% of blocks had outgroup = K-12 MG
- 62.3% of blocks had outgroup = ATCC
- 25.5% of blocks had the outgroup with a different sign than both ATCC and K-12 MG

Keep in mind that these blocks **are not** the same as the ones I am using in my analysis. So I am not sure what to do because depending on which strain is considered the "outgroup" it appears as though this ancestor is mostly similar to the K-12 MG strain or mostly similar to the ATCC strain. However, with each analysis, there are always some blocks that are in both categories (similar to MG or similar to ATCC).

Even if we did choose one of these strains, the blocks are not the same as the ones I am using in my analysis. Unfortunately, I think the correct thing to do is to do an actual reconstruction of each block (either sequence or character state) to determine what the "inverted" state should be. I found this website that discusses how to use an R package called phytools to do character state reconstruction using . This might be a quicker and simpler option, rather than doing my long reconstruction method I used in the substitutions paper.