Inversions and Gene Expression Paper Revisions:

I successfully created code to do the permutation tests for each block length (number of genes per block), and regarding both inversion patterns (checking if ATCC is different than all others and if ATCC and DH are different than all others).

What is my "observed" value in the permutation test?

I am still a bit confused by this so I am trying to write out my thoughts to see if they make sense.

Previously (before the permutation tests) I did two different Wilcoxon tests:

1. comparing expression of ALL inverted genes, to the expression of ALL non-inverted genes (lumping all blocks/genes together)
2. a Wilcoxon test on each block (homologous genes): comparing expression of inverted genes, to expression of non-inverted genes within the same block

For the second test, this means that I repeated this Wilcoxon test for each block (hundreds of blocks), obtaining multiple W statistics and p-values.

Should I be replicating these two tests/questions, but with permutations?

If so, I assume that for the first test (1. above), lets say I have a total of 1000 inverted genes and 2000 non-inverted genes, I would re-sample randomly (both inverted and non-inverted from all taxa) 1000 genes as "inverted" and 2000 genes as "non-inverted". Perform a Wilcoxon test on these samples to get a p-value. Repeat this many times. Obtain a distribution of p-values. And see where my observed value (1. above) fits along this distribution.

For the second test (2. above), I am getting a bit confused. Currently, I am splitting the data up into block length by gene. So if a block contains 3 genes it's length would be 3. I perform a permutation test on each block length. In this example, sample various columns from the alignment (homologous genes) to end up with a re-sampled block the same length as the original block (in this case 3). Compare expression of the ATCC strain vs. the other strains using a Wilcoxon test to get a p-value. Repeat many times. Obtain a distribution of p-values. And see where my observed value (from the original block) fits along this distribution. Since each block has its own observed p-value, I end up with multiple p-values. **Would I have to test each p-value from each block against the permuted distribution of p-values generated from re-sampled blocks of the same length?**

Ancestral Inversion

I was looking over the reviewers comments to see what was left to do (which is not much!), and I noticed that we did not decide (or I can't remember) what to do about how we are "defining" inversions. The reviewer is concerned that by arbirtarally picking K-12 MG1655 as the reference, we may be inaccurately identifying inversions. I.e. what if K-12 was actually inverted (and therefore all other strains), making ATCC not inverted. They repeatedly mentioned determining what the ancestral inversion is and using this as the base for determining the inversion status of each gene.

To address this, I attempted to look at various outgroup strains of *E. coli*, determine their inversion status, and see if it generally matches ATCC or K-12. If it always matched K-12, then we could confidently say that choosing K-12 as the reference for inversions was a sound choice. The results are summarized below.

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.

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 (because PARSNP re-calculates the core blocks based on what taxa are present). 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.

I am unsure of what to do next. Is ancestral reconstruction worth it? Should I just justify in the cover letter that inversions are truly arbitrary (depending on the reference), but no matter who is the reference, we see that ATCC is usually in a different state compared to the other taxa? Let me know what you think.