Inversions and Gene Expression Paper Revisions:

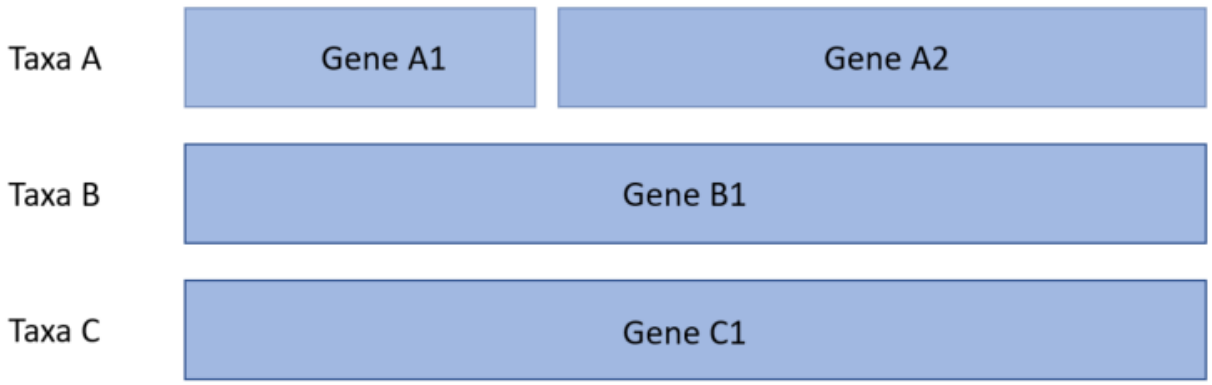
I had to alter my data frame slightly to associate the normalized expression with homologous genes. I now have a nice table where each row is the expression value of a homologous gene, and the columns are the different taxa. This will make it easy to grab rows randomly for the permutation.

What to do with overlapping genes?

I have a few cases where a gene in one taxa is large, and encompasses two (or more) genes in another taxa (Figure , please do not judge my hideous use of powerpoint. My new lab does not understand latex and they do not like automation (i.e. they use illustrator for EVERYTHING and then complain when they need to spend an hour editing it. I am working on changing them.). At the moment, I consider each of the genes in taxa A as separate homologous genes to the genes in taxa B and C. So, the sets of homologous genes would be:

- gene A1, gene B1, gene C1
- gene A2, gene B2, gene C2

Does this make sense? and, should I consider these separate "columns" for the permutation analysis? i.e. when I grab a column of homologous genes, is it ok for me to consider each of the above sets of homologous genes as a potential column?



Permutation Test

I want to ensure that I understand the test correctly. First, I create my new permuted blocks with the same number of genes as the original block length (i.e. if the block was composed of 2 genes, my permuted block will also contain 2 genes). Second, I do a test (Wilcoxon?) to determine if

the expression of the ATCC gene (which was inverted in the original block) is significantly different than the others. **Is this the test I should be doing?** Third, I repeat the first two steps many many times. Fourth, create a distribution of the p-values. Fifth, see where my original p-value falls within that distribution.

Is this general process correct?