Project proposal for programming for biologist 2017- Ying Sun

Q: What motifs regulate the expression of X genes across plant species?

Data sets:

1. Gene list: can be found on Neomorph
2. GFF file
3. Genomes for Arabidopsis, *B. rapa, S. parvula, E. salsugineum, S. irio, C. rubella*

Problem: How can we compare differences in motif position and motif sequence variation across plant species?

Approach 1: Using TF binding data
DAP-Seq datasets were generated to profile the binding sites of TFs responsive to ABA, a hormone that regulates abiotic stress in plants. This is similar to ChIP-Seq and the output of this analysis is a narrowpeak file that contains the start/stop positions of peaks called by GEM or a gene list after these peaks were associated with the nearest gene. How can we use these data to compare motif sequence and position across multiple plant species?

Approach 2: Using RNA-Seq data
An RNA-Seq dataset was generated to identify genes differentially regulated upon NaCl treatment in plants. Using this list, what motifs are upstream of these genes in the promoters?

1. Parse through each genome using GFF files to identify TSS for genes of interest
2. Identify promoters (1KB) upstream of TSS for a gene of interest
3. Do this for gene orthologs across 6 plant species
4. Format this dataset to generate an input file for MEME
5. Compare motifs in MEME
6. Identify and count variants within the promoter across species. Identify how many variants lie within the motifs across each species.
7. Output results to text file
8. Do this for many genes then group them by pattern
9. If time: Compare protein sequences for these genes by alignment
10. If time: Can discuss other methods for associating peaks from bed file (CHIP or DAP output) to genes to get a better gene list
11. If time: can discuss data visualization, how will we represent variation across 6 species?

Things we will learn:

1. Parsing/ organize/ formatting files so they are useable (could use Biopython?)
2. Getting nucleotide sequences from the genome
3. Comparing nucleotide and AA sequence by alignment
4. Learn how to count variants.
5. Getting a formatted output file for downstream analysis