Armillaria Gallica Gene Analysis Tools

Read Depth Snapshots

In order to gain a understanding of how the locaitons which Hao had determined had notably high read depth the regions Hao identified were graphed. There were approximatly between 2000 and 4000 regions for each of the 15 strains. Although there were many graphs four predomanant types of high read depth regions stood out: Normal, Rectangular, Right skewed, and Left skewed.

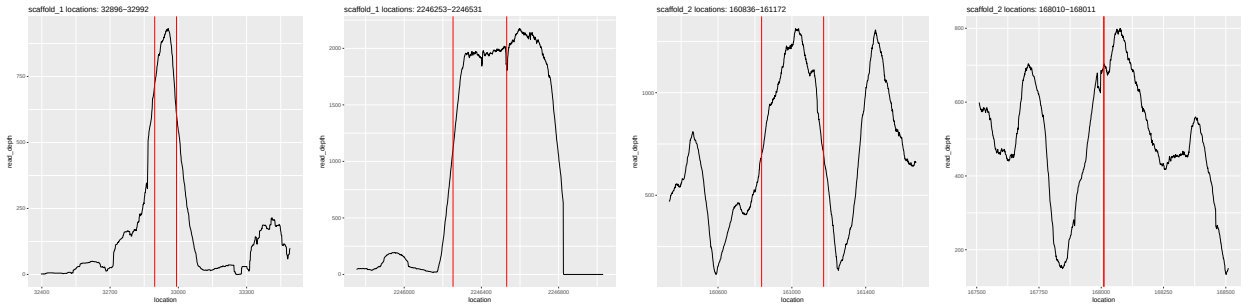

Figure 1. Examples of the four types of high read depth regions (left most) Normal, (left middle) Rectangular, (right middle) Left skewed, (right most)Right skewed

Average Read Depth

Due to the large quantity of data present in each bam or fastq file we made use of the meta data of the aligned reads in the form of the average number of aligned reads. We calcualted the average for each strain and scaffold individually, and also calculated the global average read depth. Locations that did not have any reads aligned to them were not taken into accout. The global average read depths are shown in table 1 and a example of the average read depth per scaffold and the number of aligned reads for each scaffold are graphed in

figure 2.

Table 1. Gobal Average Read Depths and Number of Reads for Each Strain

Strain	Global Average Read Depth	Total Global Number of Reads
Ar73	111.0507	69955093
Ar109	117.6863	70143802
Ar119	112.4868	70015910
Ar142	109.3741	70068064
Ar159	104.3773	69875550
Ar170	112.3987	70033946
Ar174	113.4283	70022954
Ar175	73.79959	69545937
Ar176	73.21531	69583274
Ar179	117.2196	70061598
Ar188	63.61699	68627951
Ar194	67.88522	69488072
Ar196	113.9182	70027951
Ar201	68.39596	69488227
Ar213	110.9612	69988055

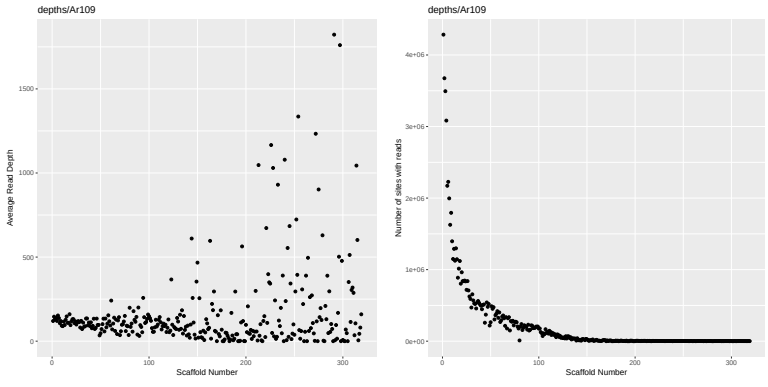

Figure 2. Average Read Depth Per scaffold and nubmer of aligned reads per scaffold, strain Ar109. (left) average read depth, (right) number of aligned reads.

Identifying Regions of High Read Depth

There were two issues with the methodology Hao used to identify regions of high read depth. The first was that if the read depth peaked over the theshold for only a few locations then it would be captured in his search. This resulted in many locations of significance ranging only a few nucleotides. The second issue was that if the region of significance dipped below the search threshold then immeadiatly rose up above then a region which should have been contigious results in two regions identified. We attempted to solve this second issue by creating out own significant read depth search program. This program searched for all locations which were five times the standard deviation above the mean read depth for that scaffold and outputted those regions. The program also allowed for a customisable *grace* peroid (set to 10 locations currently) where if the read depth at a loation drops below the threshold for less than the grace peroid then the regions is treated as contigious. An example of a read depth snapshot is shown below in figure 3.

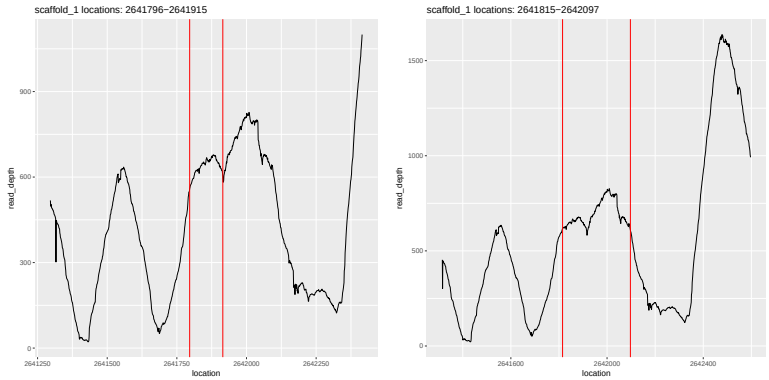

Figure 3. Significant Read Depth Identification Differeneces. (left) Hao's method breaks the hill into two parts, (right) method with grace peroid captures intact read depth hill.

Consensus of Aligned Reads

Between the identification of locations where there were indels and having many locations where significantlly high read depths were found it would be useful to know what the sequence at a region. To do this a series of programs were created which could take in a indexed bam file, a scaffold, a start location, and an end location. This program would then output the mode of all reads aligned within that region.

Graphing Read Depth Across Scaffolds

To take a birds eye view of the read depth for each scaffold, we separated each scaffold read depths into their own files and graphed the individial scaffold read depth. The scaffolds vary in size greatly, as can be seen in figure 2. Shown below in figure 4are scaffolds 1 and 200 for strain Ar109 and the histograms of the read depths.

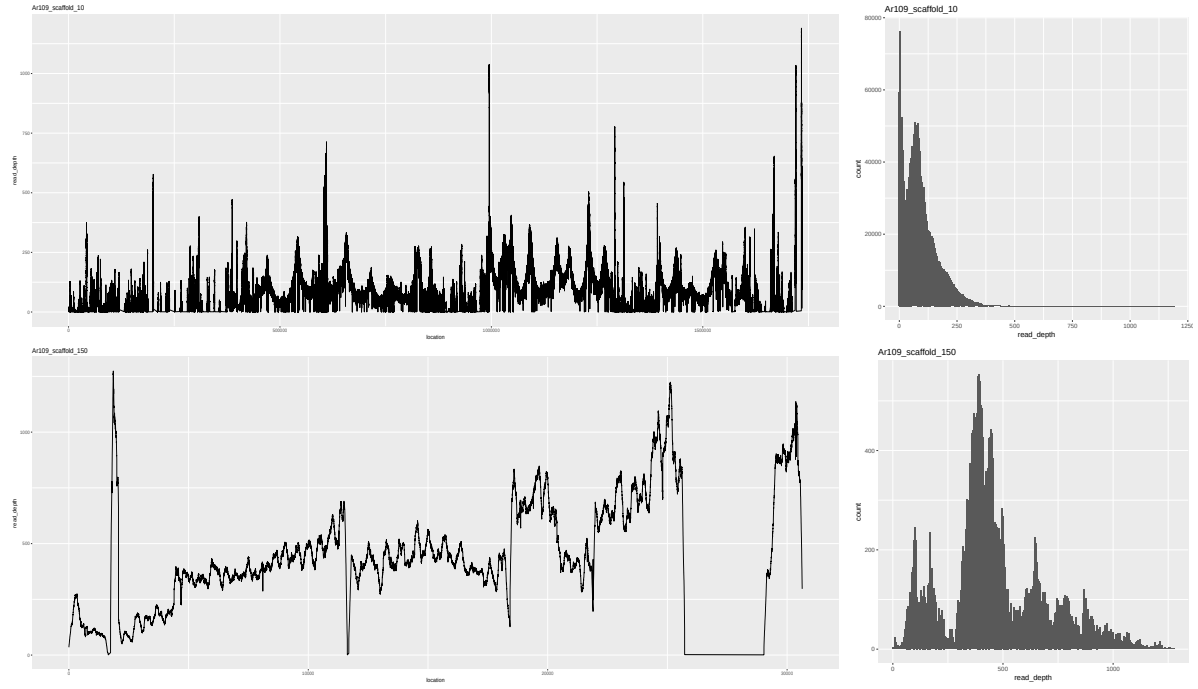

Figure 4. Whole Scaffold Read Depth Graphs for strain Ar109. (Top left) Scaffold 10 read depth, (Top right) scaffold 10 histogram, (Bottom left) Scaffold 150 read depth, (Bottom right) scaffold 150 histogram.

Read Depth Differences Between Strains

To compare the read depths between strains we created a program which would iterate over two read depth files and output the differences in read depth at all locations. Due to the similarity of the output from this program to the input used in the program to find significant read depths we were able to perform the analysis same analysis outlined in the section on identifying regions of high read depth. These sequences could then graphed and blasted.

Assemblies

Did velvet assemblies - did not work Did novoalign assemblies - worked but no time

Sequence Identification With Blastn

- searched for some stuff but didnt get a lot of hits

RNA Verification of Blast Method

Indel Analysis

Future Directions

- Attempt to find larger indels using a different methodology
- Take full inventory, via blastn, of all the indels identified and the immediate regions surrounding them
- Create a more robust method to search for regions of high read depth
- Take full inventory of all the sequences at regions of high read depth
- Look at the sequences which result from comparison of
- Search for transposons
- Complete De novo assemblies and carry out many of these analyses on those
- Look into the reads which did not align to the reference and attempt to find any variation which may exist between the strains