

Armillaria Gallica Gene Analysis Tools

Read Depth Snapshots

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

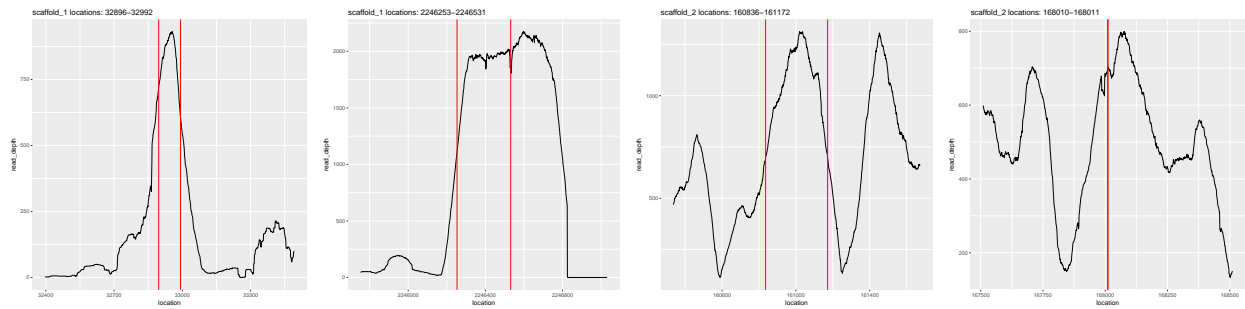


Figure 1. Examples of the four types of high read depth regions (left most) Normal, (left middle) Rectangular, (right middle) Left skewed, (right most) Right skewed

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 calculated 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 account. The global average read depths are shown in table 1 and an example of the average read depth per scaffold and the number of aligned reads for each scaffold are graphed in

figure 2.

Table 1. Global 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

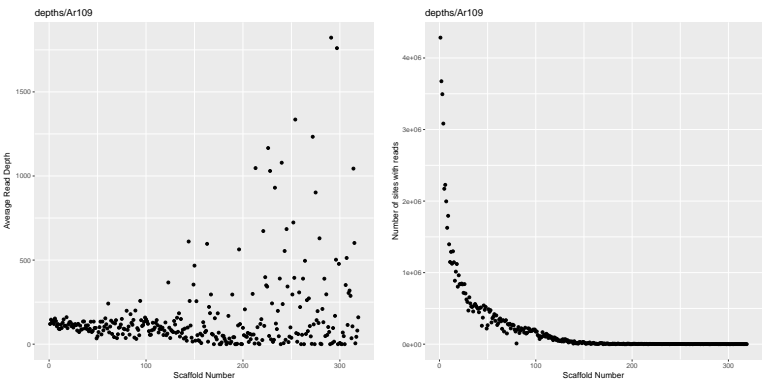


Figure 2. Average Read Depth Per scaffold and number of aligned reads per scaffold, strain Ar109. (left) average read depth, (right) number of aligned reads.

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 threshold 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 immediately rose up above then a region which should have been contiguous results in two regions identified. We attempted to solve this second issue by creating our 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* period (set to 10 locations currently) where if the read depth at a location drops below the threshold for less than the grace period then the regions is treated as contiguous. An example of a read depth snapshot is shown below in figure 3.

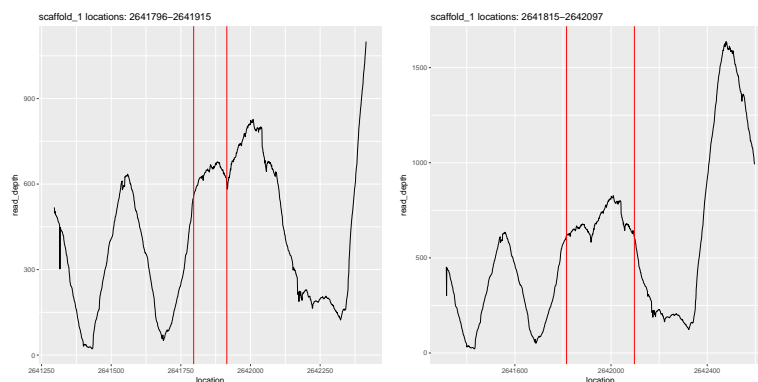


Figure 3. Significant Read Depth Identification Differences. (left) Hao's method breaks the hill into two parts, (right) method with grace period captures intact read depth hill.

Consensus of Aligned Reads

Between the identification of locations where there were indels and having many locations where significantly higher 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 an 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 individual scaffold read depth. The scaffolds vary in size greatly, as can be seen in figure 2. Shown below in figure 4 are scaffolds 1 and 200 for strain Ar109 and the histograms of the read depths.

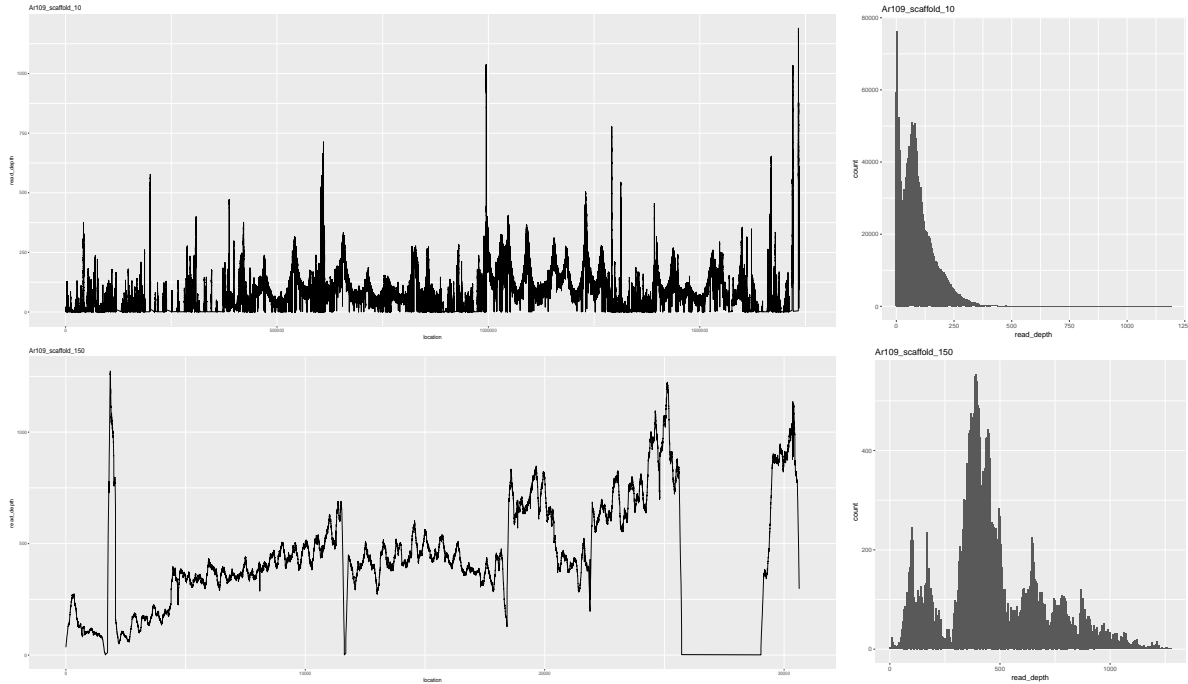


Figure 4. Whole Scaffold Read Depth Graphs for strain Ar109. (Top left) Scaffold 10 read depth, (Top right) scaffold 10 histogram, (Bottom left) Scaffold 150 read depth, (Bottom right) scaffold 150 histogram.

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

The bam files which we were able to work with were created using reference based assembly, but the unaligned reads were stripped from the file. One of the experiments which we wanted to carry out was to create an assembly of the unaligned reads and to see how those 15 assemblies may have compared to each other. In order to gain access to these reads we attempted new reference based assemblies using five of the 15 strains. Although these assemblies have been completed, they have not been analyzed yet due to time constraints.

We also attempted de novo assemblies of five of the 15 fastq sequences we had. These de novo assemblies were carried out using Velvet and VelvetOptimiser, using default settings. Unfortunately the results from these assemblies were of low quality. The largest contig was roughly 5000 bases and the total number of contigs was over 150,000. In the future we will attempt to determine why these de novo assemblies turned out so poorly.

Sequence Identification With Blastn

Making use of the identified regions of high read depth and the program which can extract sequences from the bam file we were able to use blastn to attempt to determine the function of the different notable sequences. Due to the high number of sequences we were only able to search a small fraction of the total, and the results from those searches were inconclusive. Nearly all of the sequences searched had no hits and the few that did were mainly mRNA sequences. We determined that there is a method to carry out multiple blastn searches quickly, although we have not yet been able to make the program needed for this function. When we can get this method working we will be able to carry out an automated blast search

of any sequence we find interesting.

RNA Verification of Blast Method

Due to the low number of significant blast results, we attempted to verify this method of search by attempting to find a sequence which we know to be present in *Armillaria Gallica*. Specifically we obtained a partial 18S rRNA sequence for the fungus *Aspergillus niger* that was roughly 1700 nucleotides long, and blasted that sequence against the reference genome for *Armillaria Gallica* on the joint Genome Institute Fungal Genomics Resource website. This blast returned a hit for a small region on scaffold 10 (scaffold 10:994951-995150), although this region was only 199 nucleotides in length. When we searched for this region in our scaffold snapshot graphs we were able to find that it was identified as a region of notable read depth by our system. The read depth snapshot for this range is shown in figure 5. The total ribosomal DNA for *Armillaria Gallica* should be larger than 10kb although and the 18S subunit should be larger than 3kb. We did complete a search for the entire ribosomal DNA sequence but the best we could find was a partial 28S sequence and a partial 18S/28S sequence which when blasting against the reference we found no matches. A portion of individual 28S subunit could be verified in the same method as used for the 18S subunit, but the larger 18S/28S gene could not be found. We suspect that the reference does not include the whole ribosomal DNA sequence and it may be that the ribosomal sequences present in the 15 strains of *Armillaria Gallica* are contained in the unaligned reads.

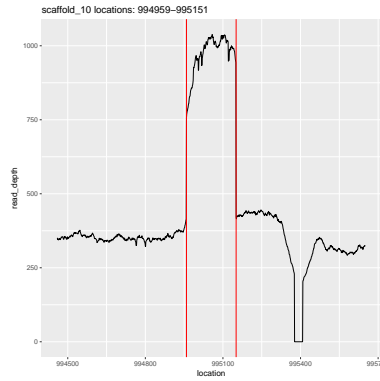


Figure 5. Read Depth for Partial 18S Sequence.

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