Armillaria Gallica Genome Analysis

Thomas Bujaki and Kassandra Dickson

Read Depth Snapshots

In order to gain an understanding of how the read depth landscape behaved, we graphed locations which previously Hao Wang had determined to have notably high read depth. There were between 2000 and 4000 regions identified for each of the 15 strains. We graphed each of these regions and all the landscapes appeared to fall into one of four predominant types of high read depth regions: Normal, Rectangular, Right skewed normal, and Left skewed normal.

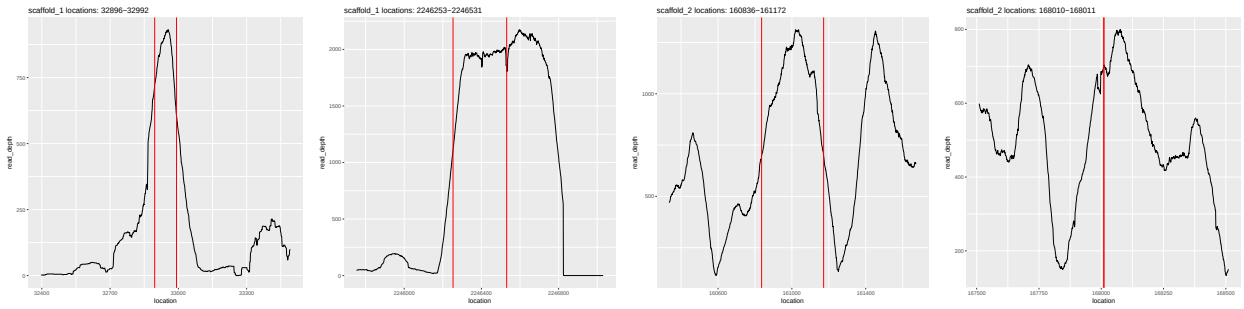

Figure 1. Examples of the four types of high read depth regions (left most) Normal, (left middle) Rectangular, (right middle) Left skewed, (right most)Right skewed

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

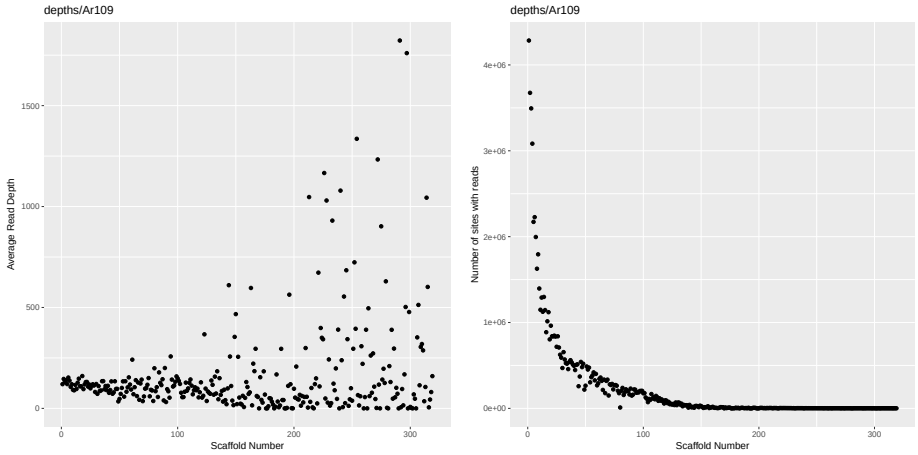

Figure 2. Average Read Depth Per scaffold and number of aligned reads per scaffold, strain Ar109. (left) average read depth, (right) number of aligned reads.

Identifying Regions of High Read Depth

There were two issues with the methodology Hao used to identify regions of high read depth. The first issue involved cases wher e the read depth peaked over the search threshold for a small number of locations, this small region would be collected as given the same weight as a larger region regardless of the significance of the short read. This resulted in many locations of apparant significance ranging only a few nucleotides. The second issue was that if a region of significance dipped below the search threshold, then immediately rose back to the level of significance, then a region which should have been contiguous results in two separate regions being 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 surrounding the dip is treated as contiguous. An example of a read depth snapshot with this issue is shown below in figure 3.

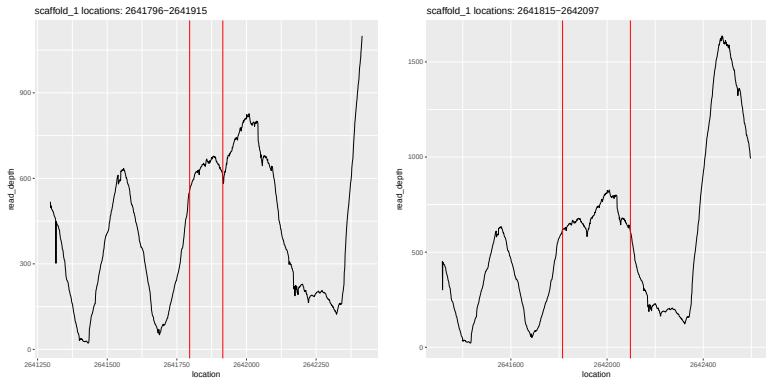

Figure 3. Significant Read Depth Identification Differences. (left) Hao's method breaks the hill into two parts, (right) method with grace period captures intact read depth hill.

Consensus of Aligned Reads

Between the identification of locations where there were indels and having many locations of significant read depth we found it nessesscary to know what the sequence at between two locations may be. 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.

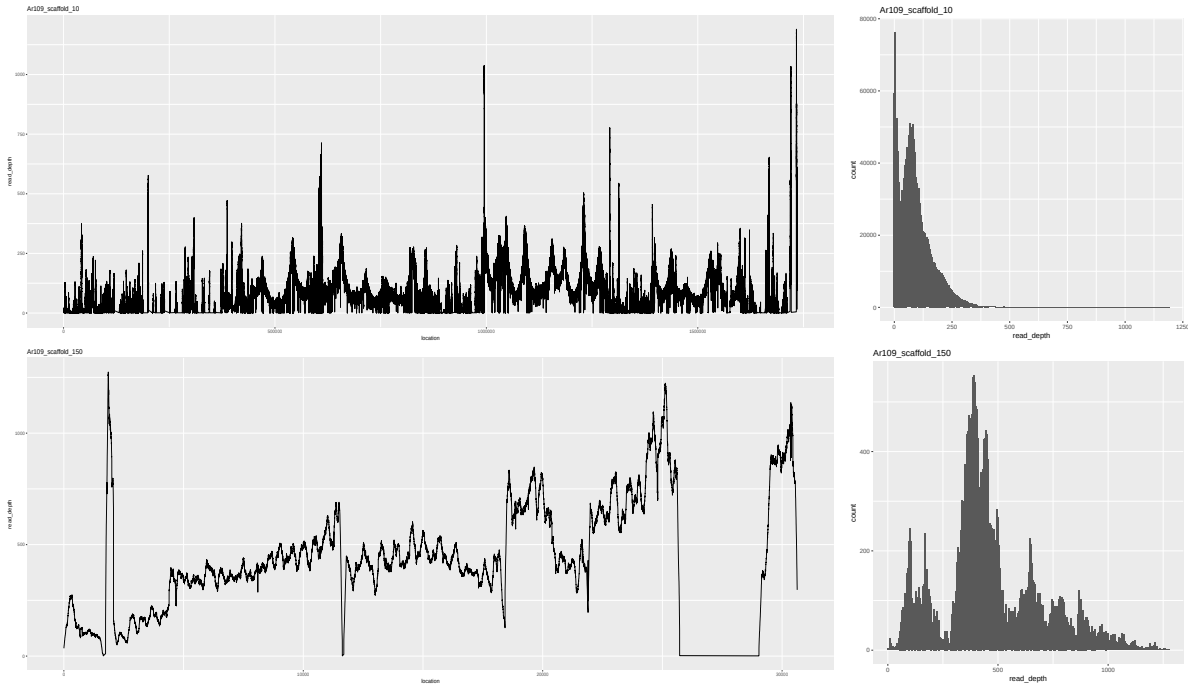

Figure 4. Whole Scaffold Read Depth Graphs for strain Ar109. (Top left) Scaffold 10 read depth, (Top right) scaffold 10 histogram, (Bottom left) Scaffold 150 read depth, (Bottom right) scaffold 150 histogram.

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 same analysis outlined in the section on identifying regions of high read depth. These sequences could then graphed and searched via 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. We also attempted combining some of the sequences into a single large fastq file. We are in the process of creating a de novo assembly of this large file.

Sequence Identification With Blastn

Making use of the identified regions of high read depth and the program which can extract sequences from bam files, 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 to 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 sequence identification 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 and the 18S subunit should be larger than 3kb. We did complete a search

on ncbi for the entire ribosomal DNA sequence but we could only find partial 28S sequences and partial 18S/28S sequences. When blasting the partial 18S/28S sequence against the reference we found no matches. A portion of partial 28S subunit could be verified using the same method used for the 18S subunit. We suspect that the reference file 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 following reference based assembly.

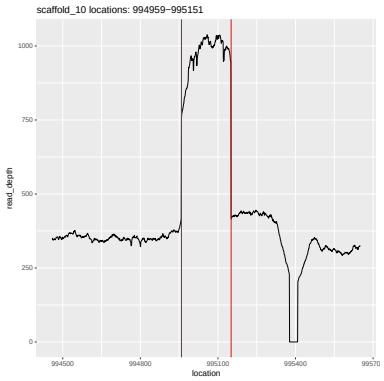

Figure 5. Read Depth for Partial 18S Sequence.

Indel Analysis

In order to start exploring InDels, a mpileup file was made for all 15 strains. This was done by using the command in samtools *mpileup*. A pileup file is a text-based format for summarizing the calls of aligned reads to the reference sequence. This format also allows for visual displays of SNP/indel calling and alignments. Once, a pileup file was produced. It is at this point, where output tags were included into the data. These output tags will indicate the read depth, quality score and also anything else that may be of use for analysis. A package called *BCFtools* was implemented. BCFtools is a set of utilities that manipulate variant calls in the variant call format (vcf) as well as its binary counterpart (bcf). The

pileup file was indexed using the command *bcftools index*. This creates an index of the sequence from the pileup file. After this is done, we use the command *bcftools call*, which is the variant calling. This will call the indels in this case and produce a .vcf and a .bcf file which show the indels and their locations. The script which does this process is called *ctq_reads_find_indels.sh*.

It appears at first glance that there was many indels detected that were the same throughout the strains as seen in 2. The summary table only shows the first three indels found within the each strain, however throughout the data, this is a repeated pattern.

Table 2. A summary table of the first three indels found in each of the strains which includes the scaffold number, the location at which the indel is found, the number of reads that support that indel, and the raw read depth

Strain No.	Scaffold No.	Location	No. of Reads Supporting	Raw Read Depth
Ar73	1	7762	54	156
	1	10784	34	148
	1	12340	37	123
Ar109	1	7762	56	163
	1	10784	68	175
	1	16154	7	176
Ar119	1	7762	62	163
	1	10784	57	140
	1	16154	4	167
Ar142	1	7762	63	189
	1	10784	55	186
	1	16154	5	125
Ar159	1	7762	41	116
	1	10784	28	100
	1	16154	3	85
Ar170	1	7762	73	222
	1	10784	61	194
	1	12340	72	193
Ar174	1	7762	63	201
	1	9593	72	218
	1	10784	45	184
Ar175	1	7762	47	141
	1	10784	28	108
	1	12340	35	102
Ar176	1	7762	39	141
	1	9593	43	129
	1	10784	33	115
Ar179	1	7762	63	193
	1	9593	64	205
	1	10784	50	195
Ar188	1	7762	17	62
	1	10784	11	47
	1	12340	24	54
Ar194	1	7762	35	133
	1	10784	32	105
	1	12340	35	110
Ar196	1	7762	72	224
	1	10784	53	169
	1	12340	72	192
Ar201	1	7762	38	110
	1	10784	44	116
	1	12340	50	102
Ar213	1	7762	76	220
	1	9593	63	196
	1	10784	48	188

Read Depth of Indels

Read depth or sequence depth was one of the output tags that we added to the pileup file previously. Read depth was important in this analysis thus, we decided to explore the them as part of the indel analysis. Since, it was already one of the output tags we previously added, it was a very simple as to extract the datapoint with the corresponding scaffold and location. These were plotted as to see which regions or scaffolds had the highest read depth and to see if the read depth overall was appropriate for further directions.

It was shown that there was good read depth generally throughout the scaffolds. The larger the scaffold, the more indels occured however, the read depth still holds to be good.

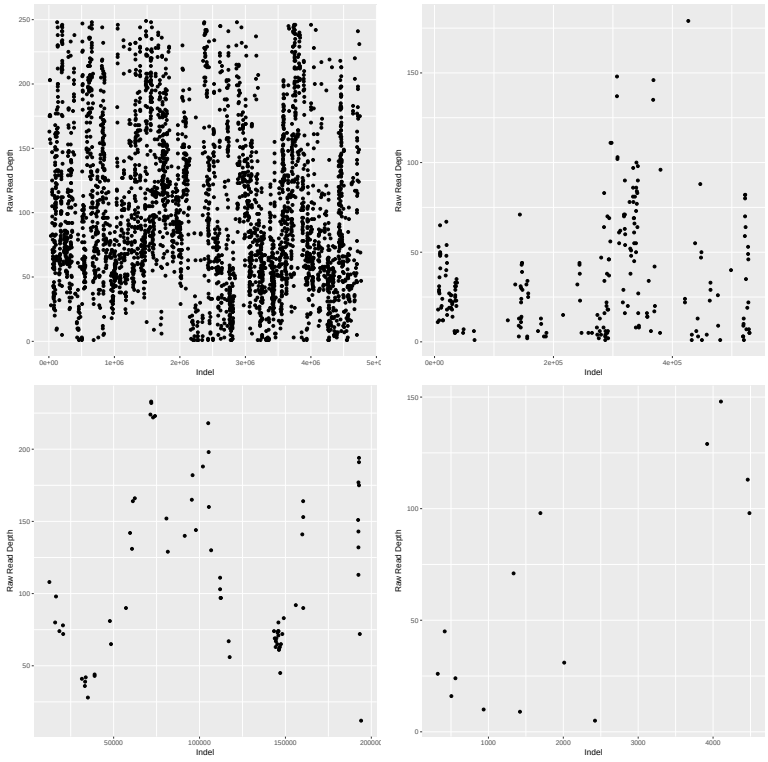

Figure 6. The Read Depth of Indels Found in Ar109 at scaffold 1, 50, 100 and 211. Every datapoint indicates one indel

Indel Lengths

Since given the read depths, it was shown that it would be appropriate to move forward. The next step we did was analyze the difference lengths of the indels. This was done by using a very simple command from the package vcftools, the command was *–hist-indel-len*. This gave a very crude histogram as well as the counts for each indel at each length. From this data we created a better visualization of the data.

It can be seen that the smaller the indel, the more frequent it occurs. This holds true throughout the strains that were tested.

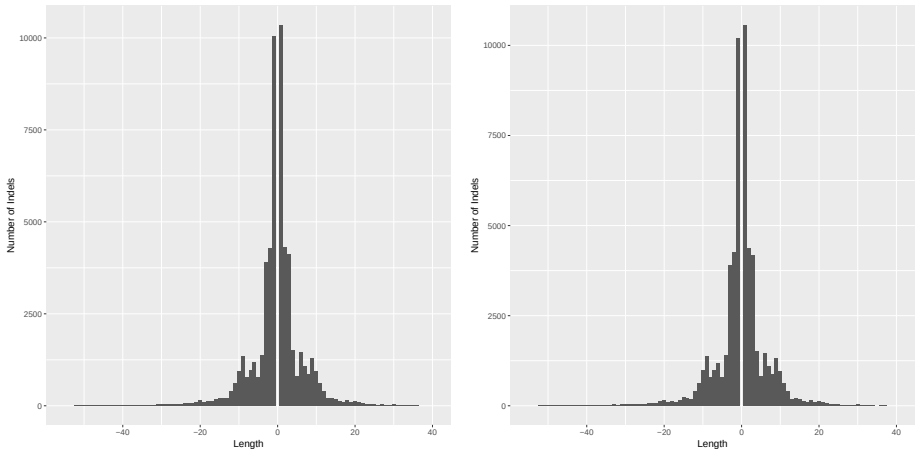

Figure 7. The Frequency of the Indel Length Per Strain, Ar109 and Ar119

In the histograms, it can be seen that the largest indels were found to be single nucleotide bases. It was a bell-shaped or a normal distrubution, which is what to be expected. From this, we focused on the extremities of the indels. We explored on the largest insertions detected. We extracted the sequences and the 5 nucleotides that flanked each end, and BLAST-ed these against the NCBI database.

Many of the BLAST searchs came back as empty however, there were a couple of hits throughout the data. We also found many repeated elements within the indel analysis. One

BLAST hit that was further looked into was at position 117162 at scaffold 98. It was 44 nucleotides long, which still is not that large in the grand scheme of indels. However, it was a perfect match against *Streptomyces actuosus*. This indel was found in 8 of the 15 strains that were tested. Using the literature, we indicated where this indel was found on the spatial map as well as the phylogeny.

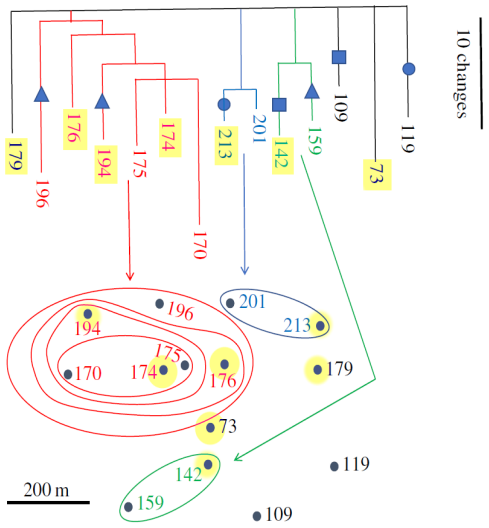

Figure 8. The phylogeny found within the original publication (Anderson *et al*, 2018) with the indication of where the *Steptomyces* Indel was found

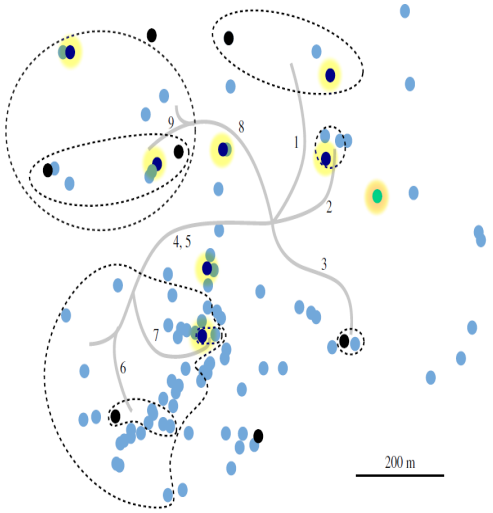

Figure 9. The spatial map found within the original publication (Anderson *et al*, 2018) with the indication of where the *Steptomyces* Indel was found

In order to show a better representation, we included the actual sequence of the found indel and the flanking regions around it. In the sequences where the indel is not present, the sequence would be linear.

Figure 10. The placement of the found indel in the sequence of the reference genome

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