Armillaria Gallica Gene Analysis Tools

Methods

This section will brief explanation of all methods used over the course of this class. For all programs discussed here see `https://github.com/ofbujak/Armillaria_gallica_gene_analysis_tools`.

Read Depth Snapshots

The first work completed was the analysis of the read depth values for each of the 15 sequenced strains of the *Armillaria gallica*. This work was based on analysis which a former student, Hao Wang, had done previously to determine locations for each strain that had unusually high read depth. These locations are based on reads aligned to a reference. Unfortunally, we do not know the exact method which Hao had used to determine which locations he had deemed had significant read depth. Though we do not know the exact method he used, the regions he identified were those that had high read depth. He outputted his results into text files consisting of three columns, a scaffold number column, a location start column, and a location end column. These files had variable numbers of locations of high significance but they averaged approximatly 2636 locations per strain. With the goal to produce graphs of these locations of high read depth, in order to gain a understainding of the "landscape" about those locations, I first made up of the samtools *depth* command in a program called make_read_depth_files.sh

make_read_depth_files.sh will output a file which has three columns, scaffold, location,

and read depth. Using these results I wrote a program, read_depth_per_location.c, which will take in as arguments a specific samtools read depth file, a scaffold to search for, a start and end location (all space separated). read_depth_per_location.c will then print information (scaffold, location, read depth, etc...) on the locations, incrementing by 1 from the start location, for all locations which are within the range passed into it. This program was used in plot_all_read_depths.sh to produce graphs of each range of locations with high read depth $\pm$ 500. The program used to create the graphs was plot_read_depth.R. Some examples of these graphs can be found in figures **??** and **??**.

Read Depths For Whole Scaffolds

In order to produce depictions of the read depths for an entire scaffold the framework established to create the read depth snapshots was used with some slight changes in methodology. To produce these graphs the files produces from using the samtools *depth* command were searched via grep in order to separate each scaffold result into its own file. These files were then passed into plot_scaffold_read_depth.R and plot_scaffold_read_depth_histogram.R. plot_scaffold_read_depth.R is a program which graphs the read depth on the y axis and the location of that read depth on the x axis. plot_scaffold_read_depth_histogram.R creates a histogram of the read depths with binwidths of 5.

Read Depth Analyses

In order to gain insight into the different aspects of the scaffolds and the reads which were aligned I produced four metrics for each strain. I calculated the global average read

depths, determined the total number of sites with reads aligned to them, calculated and graphed the average number of reads per scaffold, and graphed and determined the number of locations with reads aligned for each scaffold.

Unaligned Reads

To output all reads which are not aligned into a sam file the following samtools view command can be used: samtools view -f 4 file.bam > unmapped.sam. Following this The reads can be assembled via a de novo assembler. (The assembly of these has not been carried out yet).

New Assemblies

To produce new assemblies I attempted to do both *de novo* assemblies of five of the strains (Ar170, Ar174, Ar179, Ar196, Ar213). Following verification of quality using FastQC, both methods were used.

De Novo Assemblies

To produce new *de novo* assemblies of five of the strains (Ar170, Ar174, Ar179, Ar196, Ar213) were attempted. The program used for the assmeblies was Velvet, with the wrapper VelvetOptimiser used. The command used to run VelvetOptimiser was:

```
'perl VelvetOptimiser.pl -f "-shortPaired -fastq -separate
<Path To Forward Strand File>/<Forward_strand.fastq.gz>
<Path To Reverse Strand File>/<Reverse_strand.fastq.gz>" -t 32 -s 31 -e 31
```

```
-d <Output Directory>
```

Reference Based Assemblies

'To produce another reference based assembly the scipt fastqThroughNovoalign.sh was used. Due to time constraints, the reference based assemblies have not been analyzed as of yet.

Taking the Consensus of Aligned Read

Following identification of locations where read depth is significant it was useful to be able to quickly get the actual DNA sequence which is aligned to a specific location. To do this, custom programs were created to sort through the output from samtools view and to take the consensus of the reads aligned. The total pipeline for these programs is encapsulated within consensus_reads.sh.

To output sequences that were easier to work with, samtools view was used. The command used was:

```
samtools view -o <output_filename> <bam_file>
<scaffold_number>:<location_start>-<location_end>
```

Following outputting of the sam file, the program 'pull_reads.c' is used. This program takes in the five columns listed above (in the order discussed) and outputs all the reads, which fall within the range specified, into a file where each read occupies its own line with each read offset by the difference between the start of the range of interest and the location which the read starts.

The output created by pull_reads.c is then passed into average_reads.c. average_reads.c takes in the output created by pull_reads.c and the difference between the start and end locations. This program creates 4 arrays of the size of the difference passed into the program and iterates over all lines in the file. When a base is encounteres at a location the array for that base is incremented by 1, at the location of that base. After all the lines in the file are iterated over, the mode of each base for each element of the array are then output, resulting in a consensus of the reads that were aligned between two locations.

Calculation of Average Read Depth and Number of Aligned Reads per Scaffold

The program avg_read_depth.c takes in a read depth file and the scaffold which the average of which will be calculated and outputs the average read depths for the locations with reads aligned to them and the total number of locations with reads aligned to them for that scaffold.

Determiation of Regions Of High Read Depth

To create a new method for deremining the regions of the assembly which had significant read depth high_read_depth_regions.c was created. This program takes in the average number of reads aligned to a scaffold, and the number of locations on a scaffold with reads aligned, and searches for all locations with read depth higher than a certian threshold. The threshold used for these analyses was a multiple of five fold the standard deviation plus the mean. Any regions which were found to be above that were identified as having significant read depth.

In addition to the search this program also allows for a 10 base grace peroid. If a region was found to have significant read depth had some locations immediatly following it which were below the threshold for less than 10 consecutive locations before the read depth rose above the level of significance than the entire region, including the up to 10 bases below that region would be reported as once contigious region of high read depth.

Compairisons of Read Depths

The program compare_read_depths.c takes in two read depth files, a reference file and a compairison file. This program iterates over all the lines in the file and comapires each location and calculates the difference between read depths.

Analysis of DNA Sequences Suspected to Result in High Read Depths

Using the locations found to have significant read depth and the method used to take the consensus of the the aligned reads between locations, the sequences at each location which had high read depth could be extracted. After these sequences were extracted they could be searched, via ncbi blastn, to attempt to determine if the sequence has any homologous in other species/ to determine the effect of the sequence.

Insertions and Deletions

Kassandras stuff here.

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.

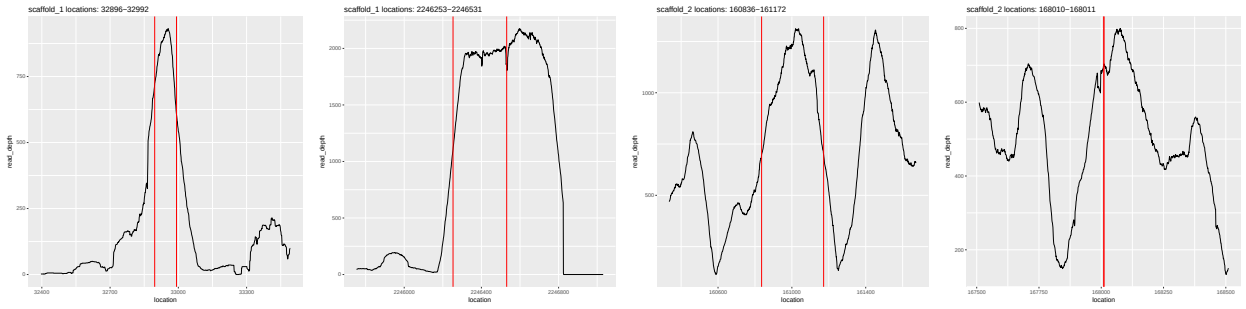

Figure 1. Examples of the four types of high read depth regions (left most) Normal, (left middle) Rectangular, (right middle) Left skewed, (right most)Right skewed

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

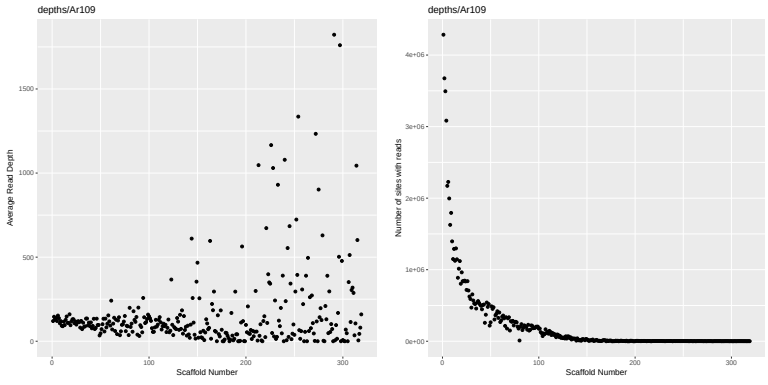

Figure 2. Average Read Depth Per scaffold and nubmer of aligned reads per scaffold, strain Ar109. (left) average read depth, (right) number of aligned reads.

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

Read Depth Analyses

Unaligned Reads

Indel Analysis

Future Directions