Armillaria Gallica Gene Analysis Tools

Table of Contents

Table of Contents i

List of Figures iii

List of Tables iv

Methods 1

Read Depth Snapshots 1

Read Depths For Whole Scaffolds 2

Read Depth Analyses 2

Unaligned Reads 3

New Assemblies 3

De Novo Assemblies 3

Reference Based Assemblies 4

Taking the Consensus of Aligned Read 4

Calculation of Average Read Depth and Number of Aligned Reads per Scaffold . . 5

Determiation of Regions Of High Read Depth 5

Compairisons of Read Depths 6

Analysis of DNA Sequences Suspected to Result in High Read Depths 6

Insertions and Deletions 6

Read Depth Snapshots **7**

Read Depth Analyses **11**

Unaligned Reads **18**

Indel Analysis **20**

Future Directions **27**

List of Figures

1 Average Read Depth Snapshots 8

2 Average Read Depth Snapshots 9

3 Global Scaffold Read Depth Example 10

4 Average Read Depth Per scaffold, Ar:73, 109, 119 13

5 Average Read Depth Per scaffold, Ar: 142, 159, 170 14

6 Average Read Depth Per scaffold, Ar:174, 175, 176 15

7 Average Read Depth Per scaffold, Ar:179, 188, 194 16

8 Average Read Depth Per scaffold, Ar:196, 201, 213 17

9 The Frequency of the Indel Length 23

10 The Frequency of the Indel Length2 24

11 The Frequency of the Indel Length3 25

12 The Read Depth of Indels Found 26

List of Tables

1 Total number of locations of high read depths identified by Hao 7

2 Gobal Average Read Depths and Number of Reads for Each Strain 11

3 Summary Indel Table 21

Methods

This section will be a in-depth 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 1 and 2.

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 attempt to gain a understanding of the read depth landscape that exists in the fifteen strains "snapshots" of the regions where it had been identified by Hao that there were notably high read depths. These snapshots consist of the region identified as having high read depth as well as the regions ±500 points from the high read depth bounds. Below are a few examples of the regions with high read depth for one of the strains, only a few of the snapshots could be displayed because Hao identified upwards of 4000 regions with high read depth for each strain.

Table 1. Total number of locations of high read depths identified by Hao

Strain	Num Significant Read depths
Ar73	2248
Ar109	4392
Ar119	3867
Ar142	2796
Ar159	3343
Ar170	2044
Ar174	1964
Ar175	2245
Ar176	2278
Ar179	2172
Ar188	3242
Ar194	2397
Ar196	2314
Ar201	2233
Ar213	2000

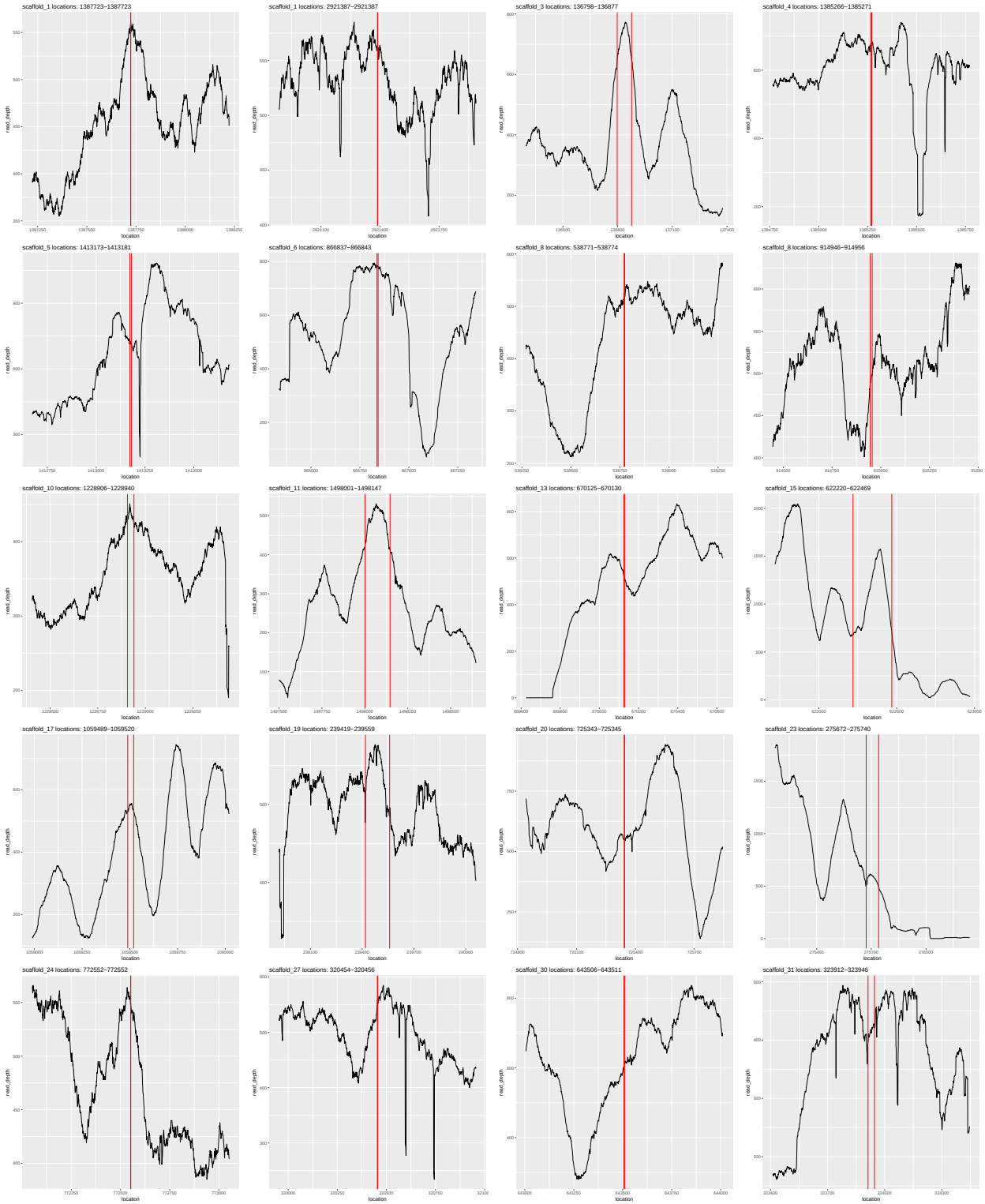

Figure 1. Average Read Depth Snapshots. Graphs are of locations 1, and every 100 starting at 100 until 1900.

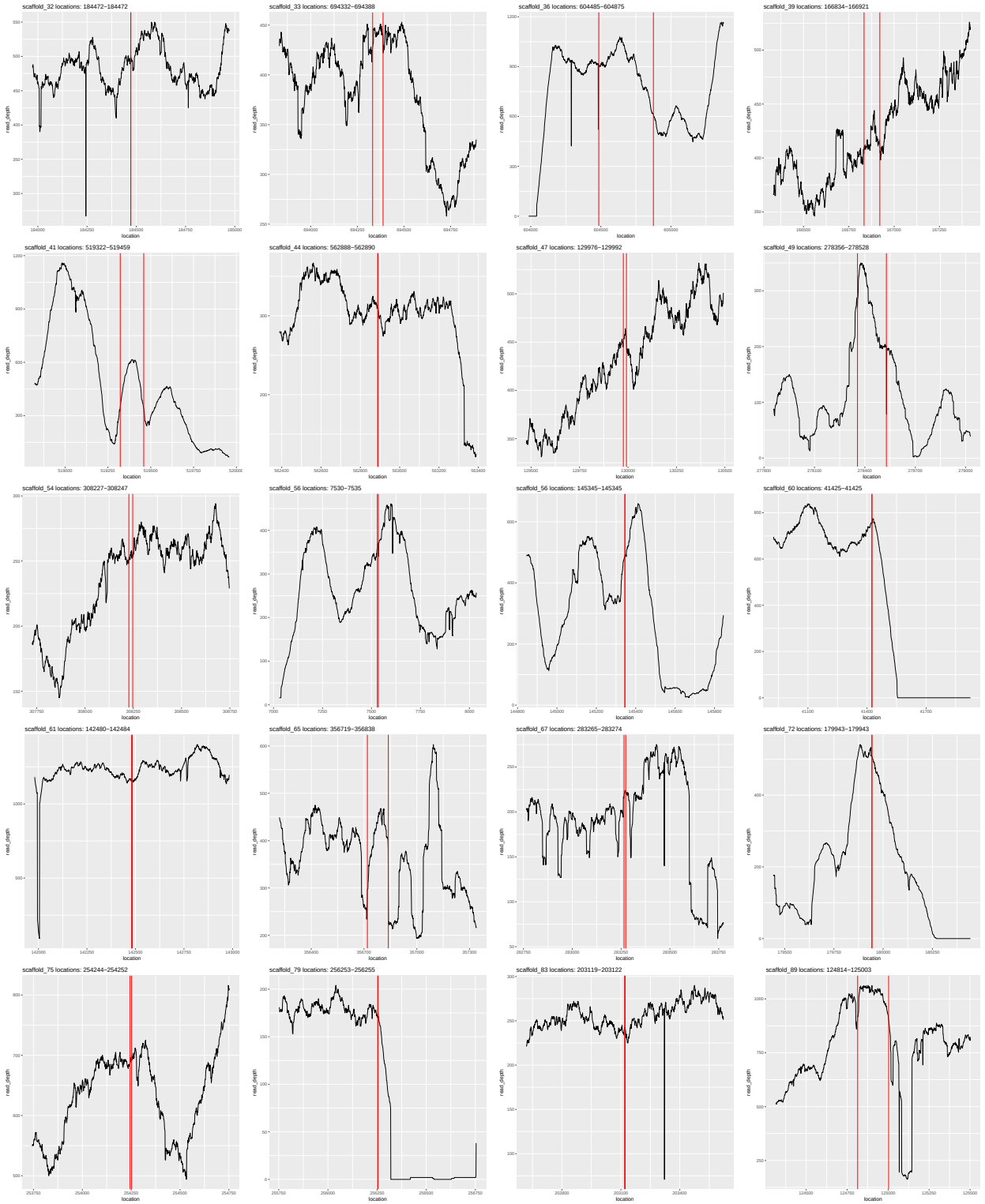

Figure 2. Average Read Depth Snapshots. Graphs are of locations 1, and every 100 starting at 2000 until 3900.

Additionally I have created a graph showing the total read depth across each scaffold. Example below.

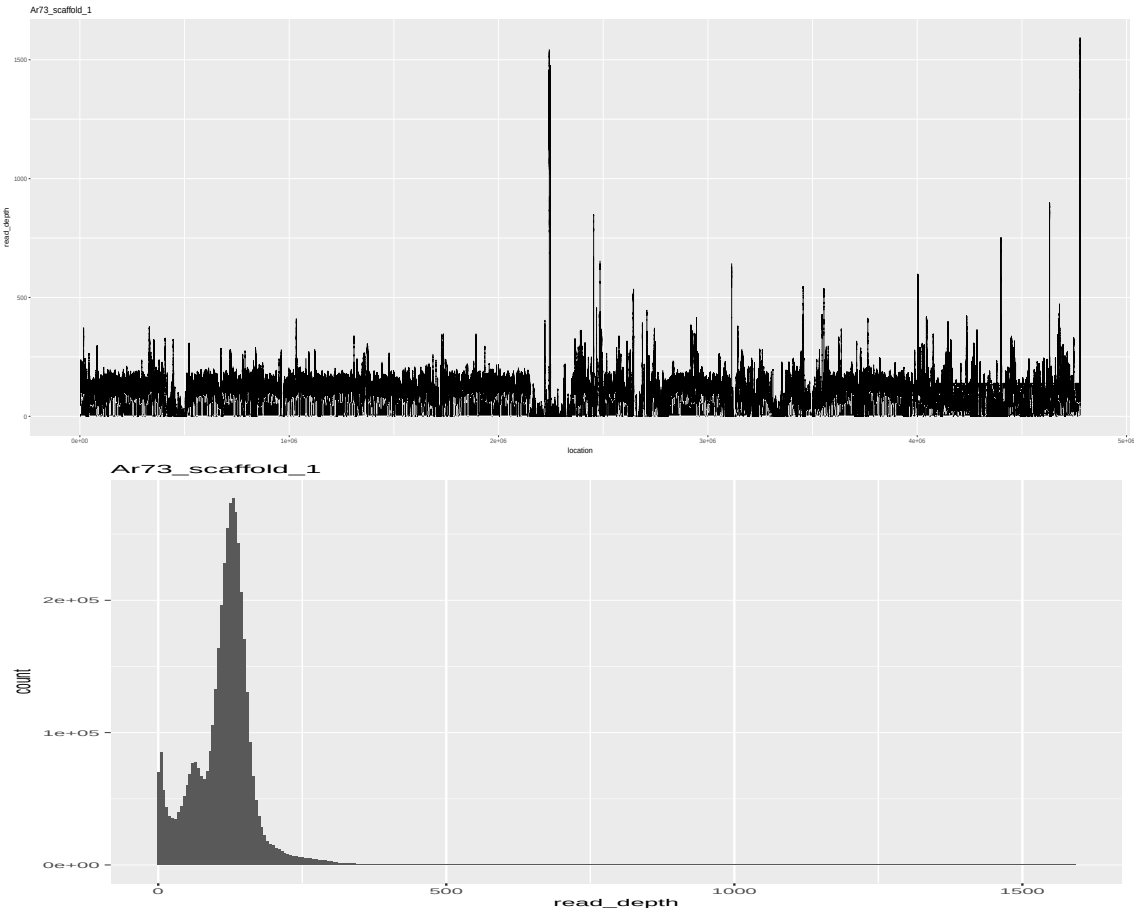

Figure 3. Global Scaffold Read Depth Example for Ar73, scaffold 1.

Read Depth Analyses

For all 15 strains of the *Armillaria gallica* fungus the global average read depths and the total number of sites, based on the reference genome, are shown in table 2.

Table 2. 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

Table 2 shows that all fifteen strains which we can work with have a similar number of sites which reads were aligned to. This number of aligned reads was between 68.6 million and 70.1 million for all strains. The global average read depths (average read depth of all sites where reads were aligned) varied much more. The low end for global average read depths was for the strain Ar188 which had an average of 63.6 reads, when reads were aligned, and the high end for average read depth was 117.7 for the strain Ar109. Hao had mentioned that Ar 188 was not used in his analyses because it was different in some way to the other strains. Ar188 is the strain with the lowest number of reads and lowest total number of reads aligned.

The average number of reads within strains was also calculated, these results can be found in figures 4, 5, 6, 7, and 8.

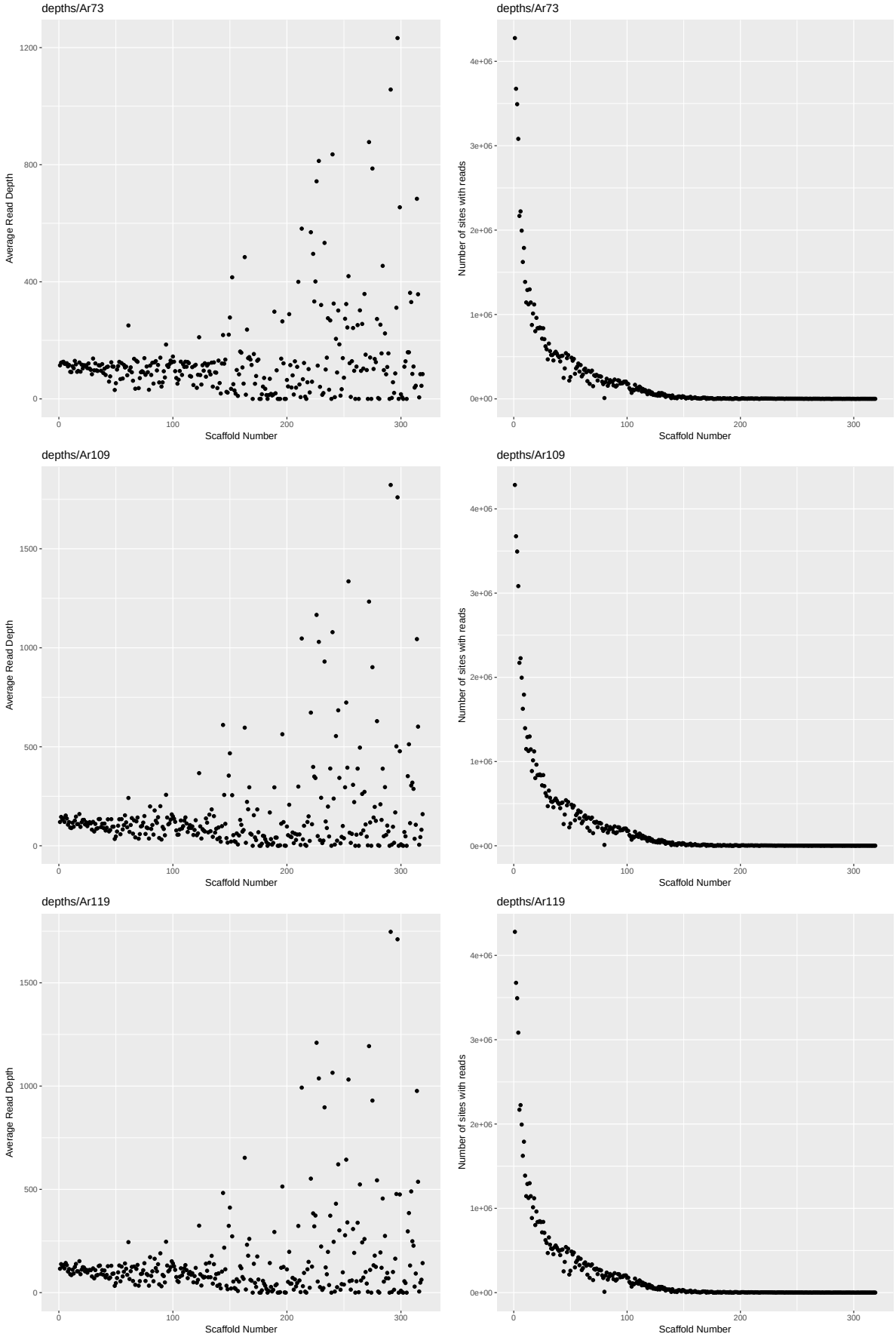

Figure 4. Average Read Depth Per scaffold, Ar:73, 109, 119. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

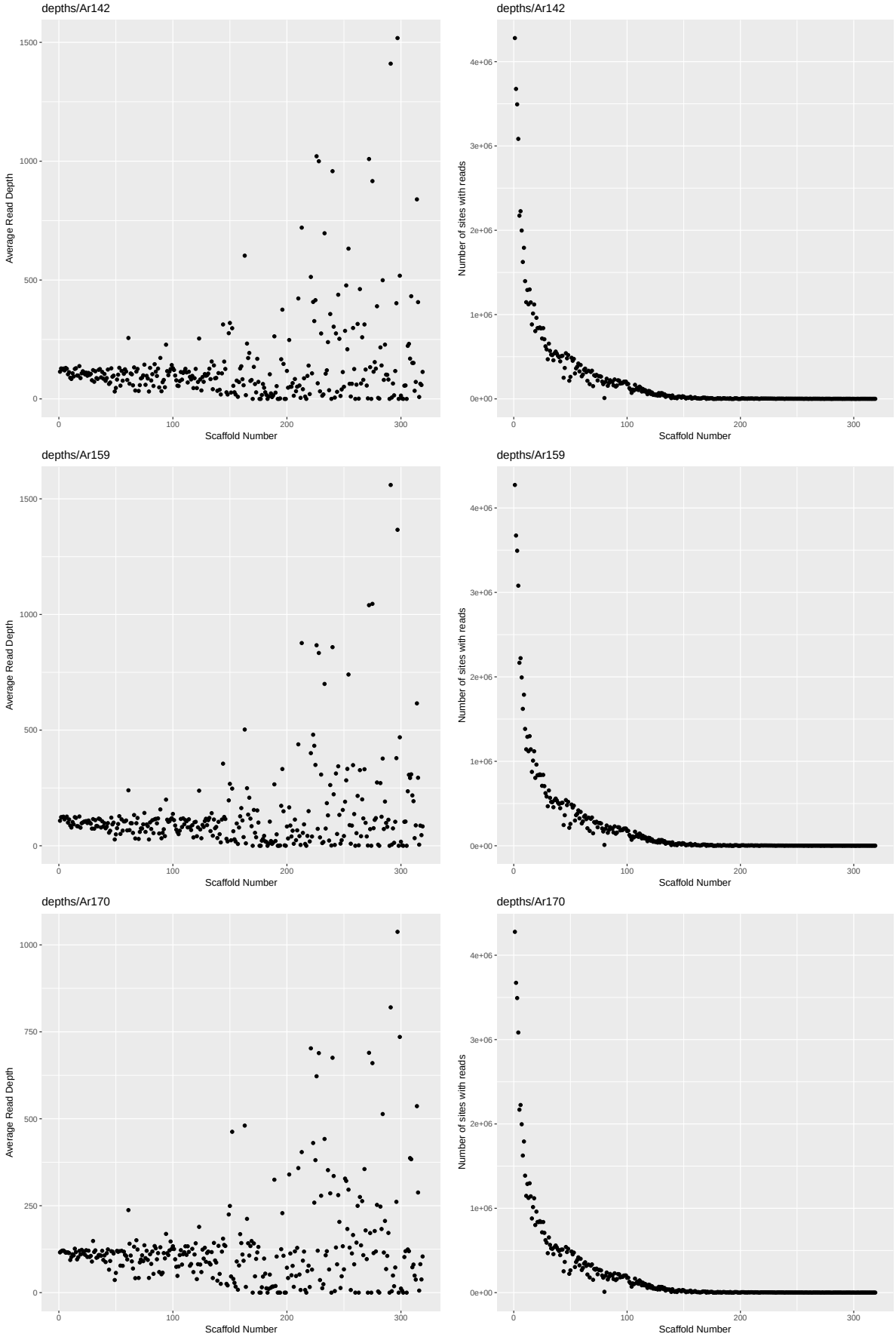

Figure 5. Average Read Depth Per scaffold, Ar: 142, 159, 170. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

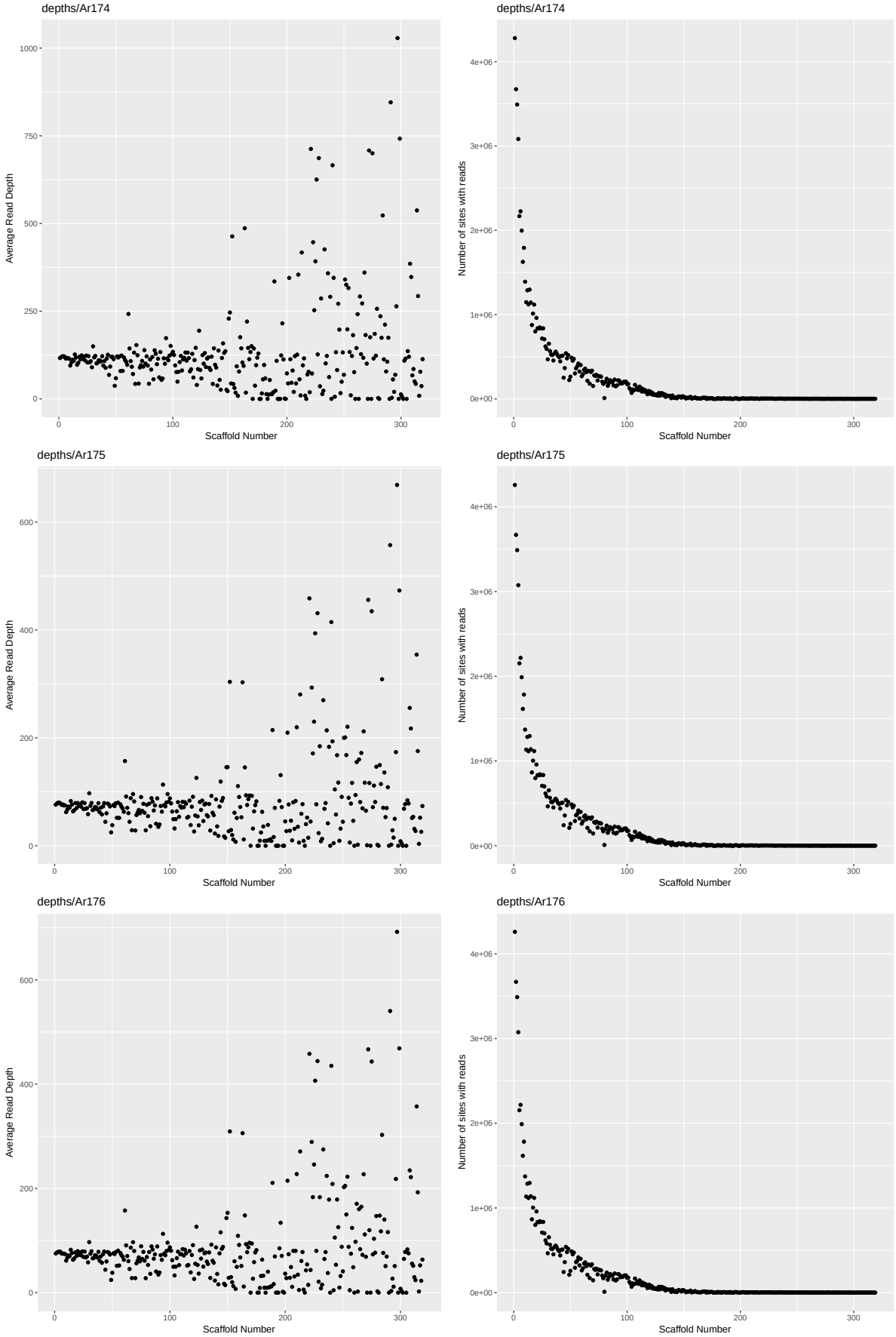

Figure 6. Average Read Depth Per scaffold, Ar:174, 175, 176. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

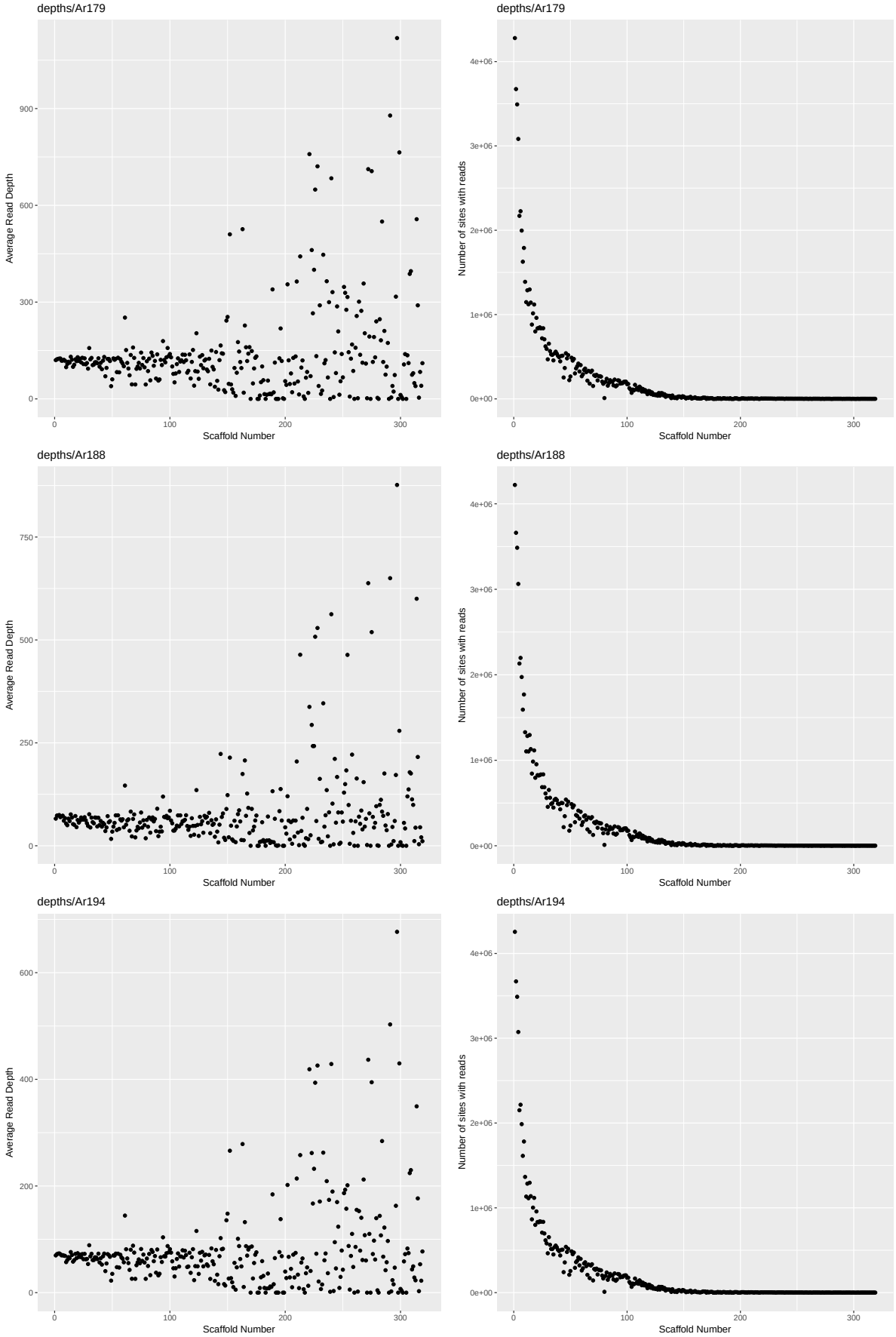

Figure 7. Average Read Depth Per scaffold, Ar:179, 188, 194. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

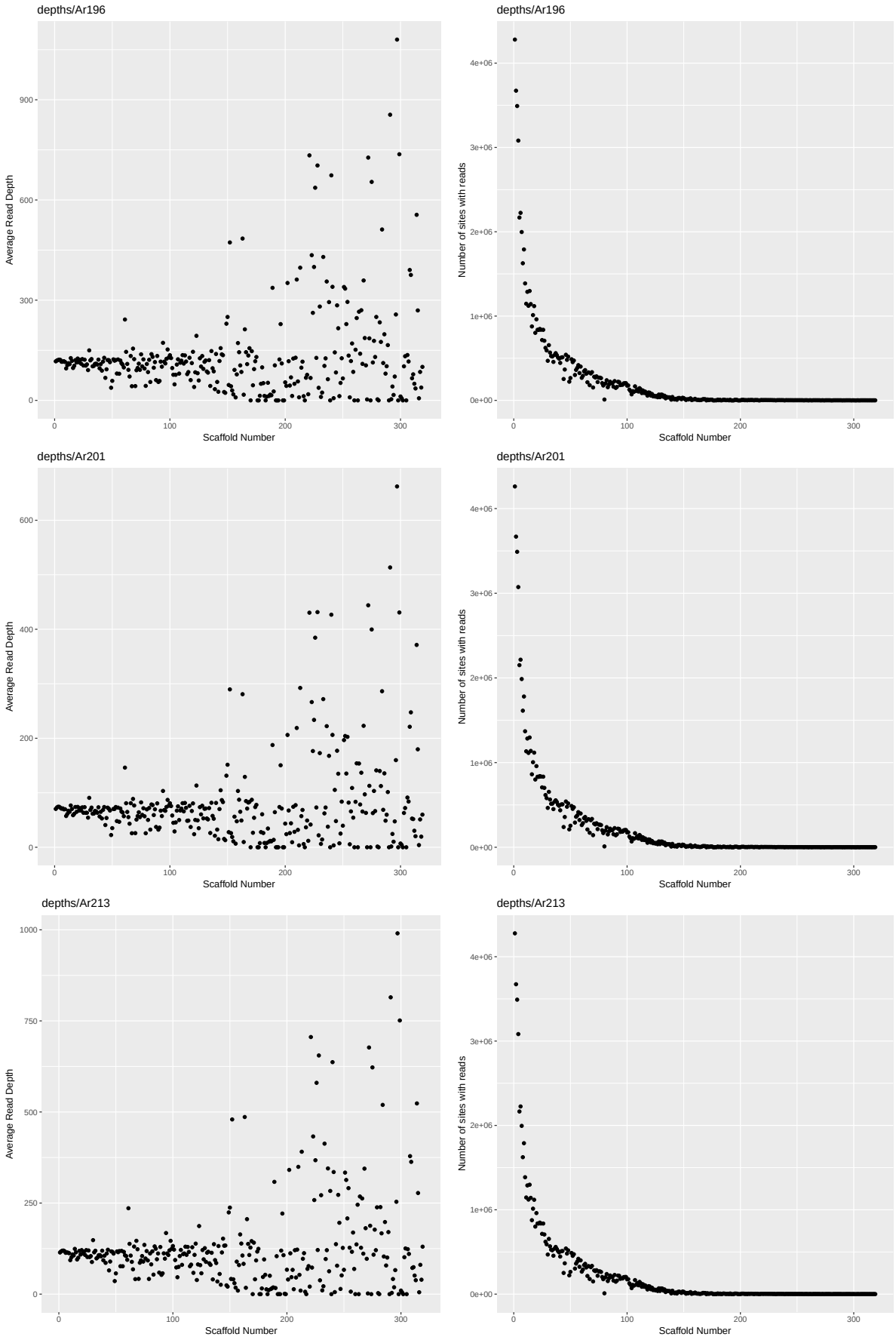

Figure 8. Average Read Depth Per scaffold, Ar:196, 201, 213. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

Unaligned Reads

It is possible that there are a number of reads which were so different from the reference sequences that they did not align onto the reference when Hao was working with them. This possibilty could mean that there would be reads which are from sequences in the big fungus that do not exist in the reference fungus genome. If this was the case than attempting a de novo alignment of these sequences could prove useful.

To see if there were unaligned reads I made use of a samtools command "samtools view -f 4 bamfile > out.sam".

But there did not appear to be any reads which were not aligned.

August 9:

I have found the original fastq sequences. After verifying that the sequences are all of high quality I have begun working on alignment of these sequences. I am working on a denovo alignment of them using Velvet and VelvetOptimisor, although becasue of the size of these sequences I need to use out labs computing cluster (I left the programs to run overnight one night and had nothing to show for it, I suspect due to RAM limitations).

I also finished the work on the program which will parse a .sam file to get sequences between locations in a aligned sequence. I used this to pull all the sequences from all locations which Hao had determined were unusually high read depth. After picking one of the strains and attempting to do NCBI blastn (nucleotide blast) on about half of the sequences which had over 100 bases (many of the regions which Hao had identified as having high reads depth are extremely short sequences, some even being only 1 base long). I found that almost all of the sequences which I searched for had no hits, and the few that did have

hits were only hits on mRNA sequences in random species.

Indel Analysis

Indels in *Armillaria gallica*

Through previous studies on the *Armillaria gallica* fungus, several strains were sequenced in Illuminia. These strains were analyzed first using samtools, in order to generate a pileup file. A pileup file is a format used to summarize the bases of the aligned reads to the reference sequence. This acts a good visual display of SNP/INDEL calls and alignment. The pileup file generated here, used labels which would indicate: maximum number of reads supporting an INDEL, raw read depth, and the number of reads that support that INDEL. Once the pileup file had been generated with the appropriate labels and filters for INDEL specific calling, a software package called bcftools was than used.

BCFtools is a set of utilites that manipulate variant calls in the variant call format (VCF) and its binary counterpart (BCF). Through this, VCF and BCF files were generated which contained genotype frequencies from each of the strains .bam files.

As seen in 3, many indels are found throughout most if not all of the strains analyzed. The summary table only shows the first three indels found within the each strain, however throughout the data, this is a repeated pattern. Indels are shown to be shared amongst the strains.

Table 3. 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

From the initial analysis of the indel locations, we decided to examine the length more in depth. We looked at the length of the indel because there was single nucleotide insertations and deletions that were found. The larger the indel, the more interest it is because these are considered less likely to be because of random insertations or deletions. A histogram was made to show the frequency found within the files of the length of the observed indels. This was done by vcftools, in which we used the command "–hist-indel-len". This command was able to generate from the vcf files previously made, the length and the frequency of the indels for each strain. 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 these histograms, we decided to explore the extremities of the normal distrubution, with the largest indels found within in the algniments. It was found the largest deletion was approximately 52 nucleotides long and the largest insertion was approximately 45 nucleotides long.

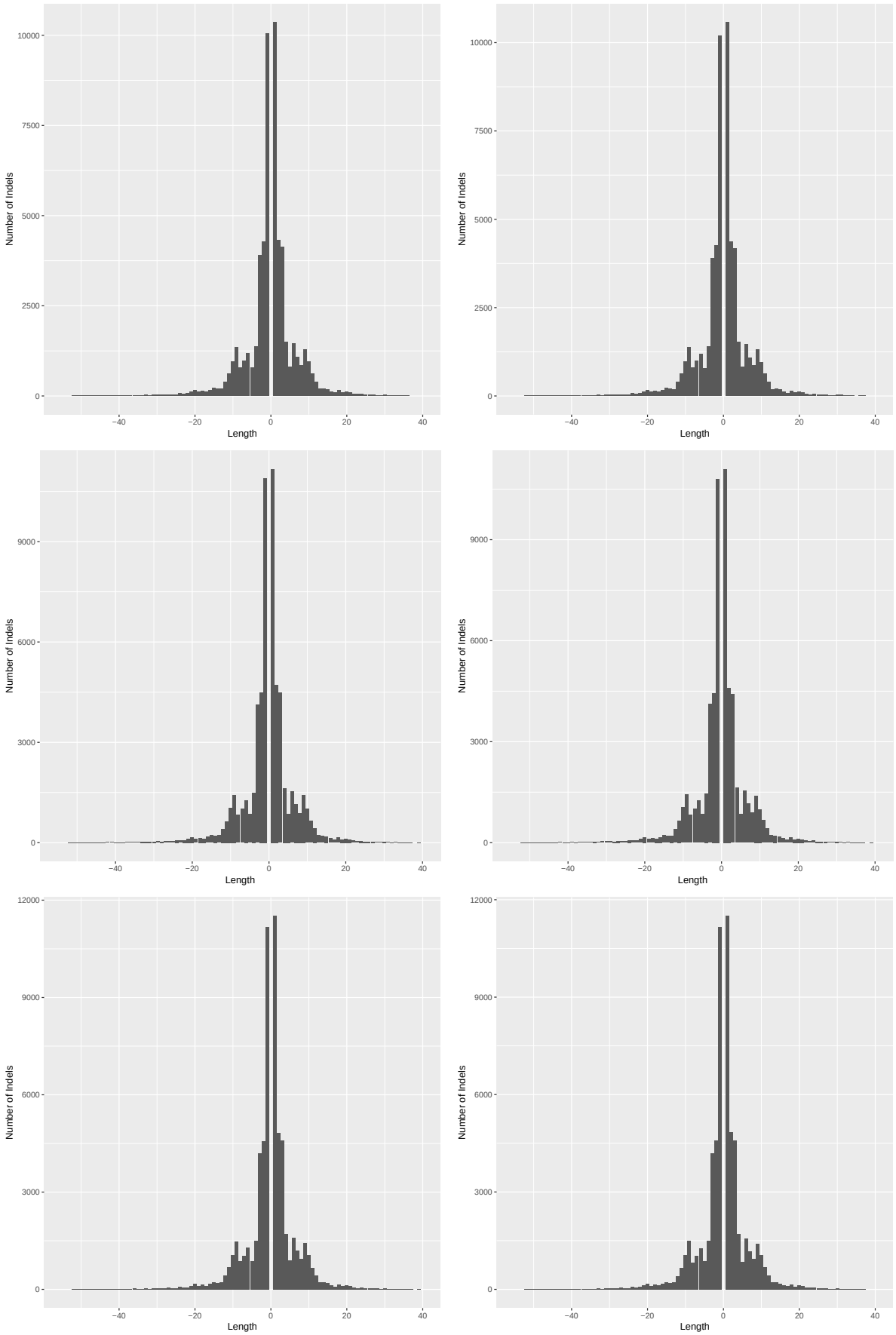

Figure 9. The Frequency of the Indel Length Per Strain, Ar: 109, 119, 142, 159, 170, and 174

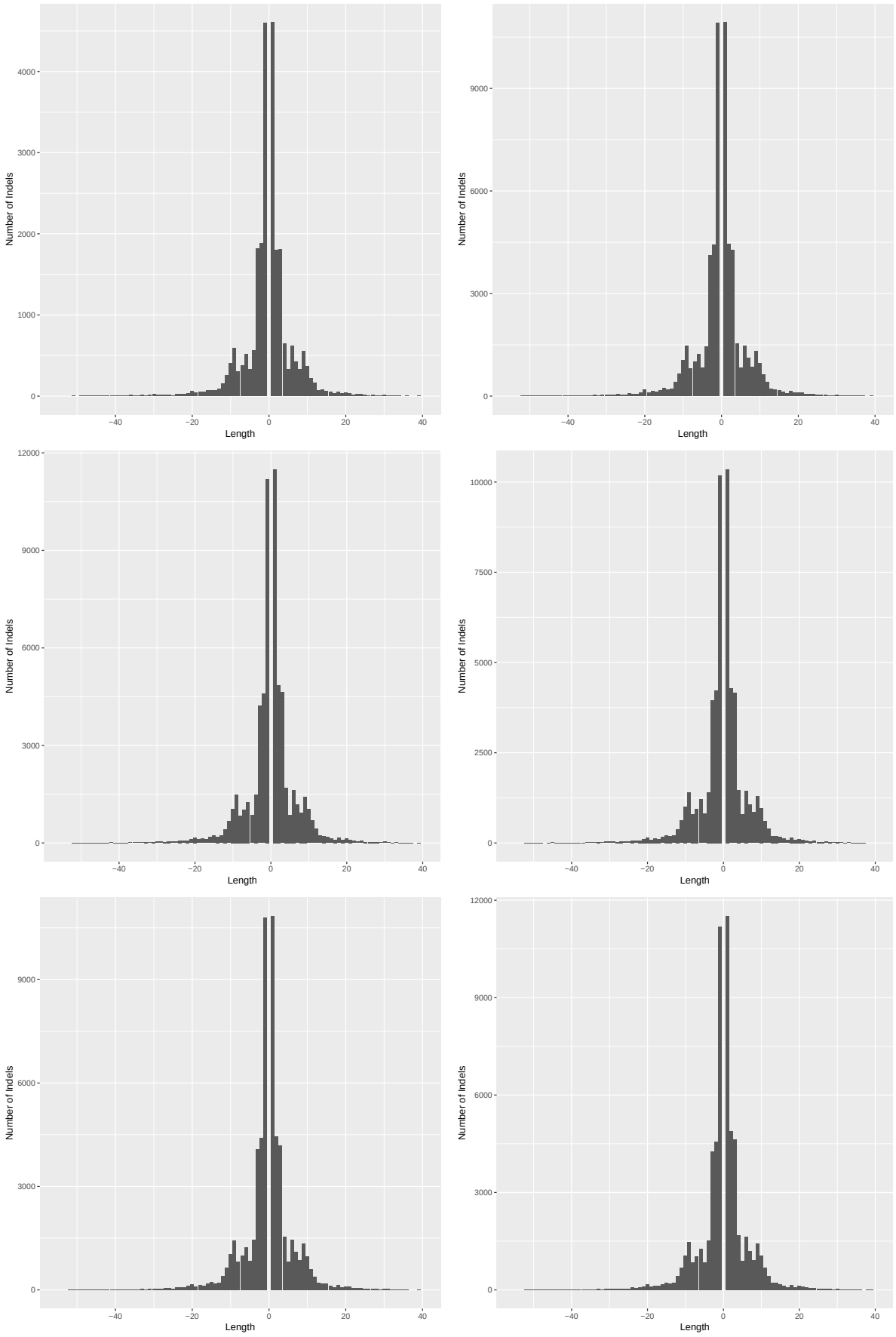

Figure 10. The Frequency of the Indel Length Per Strain, Ar: 175, 176, 179, 188, 194, and 196

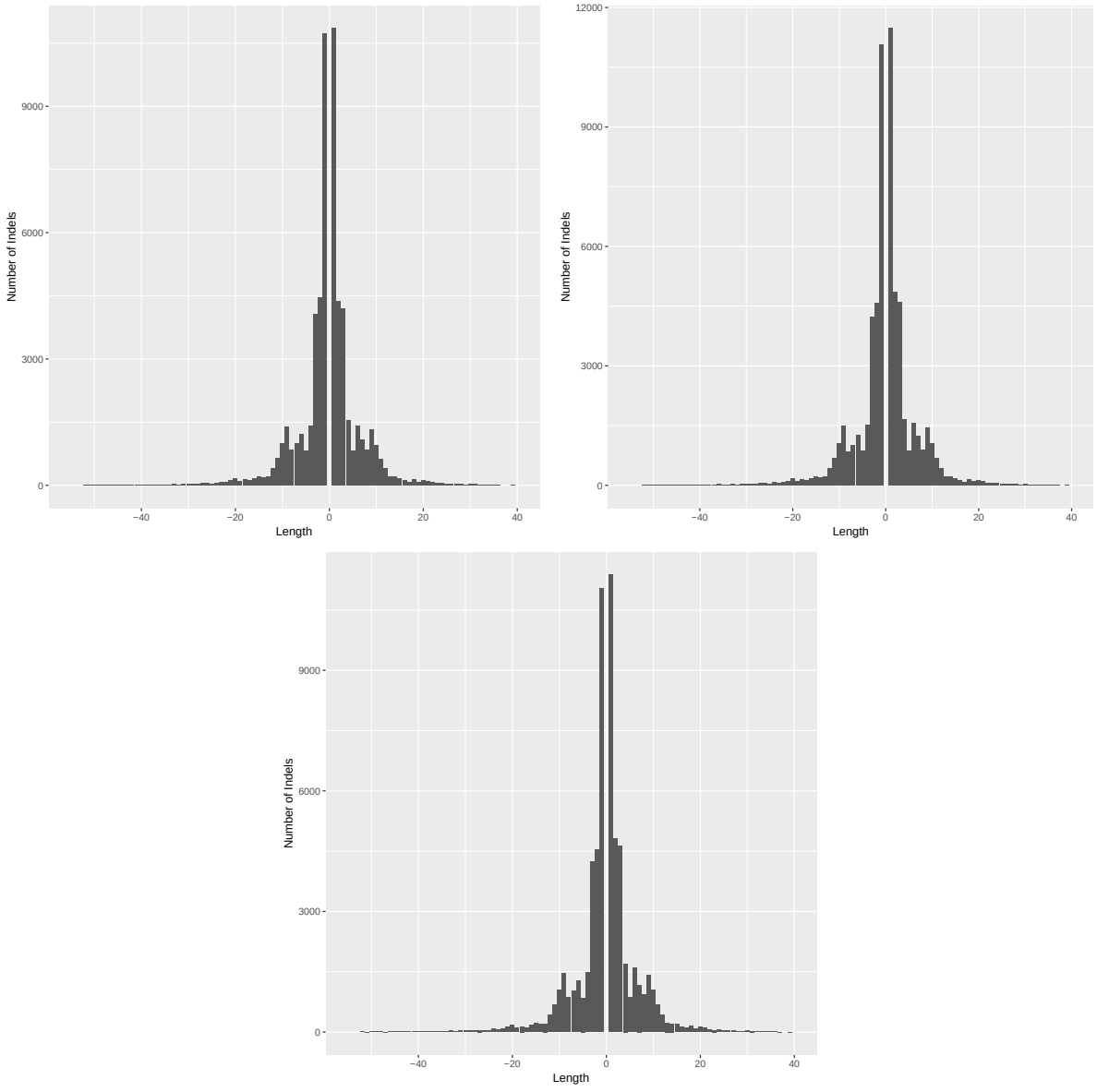

Figure 11. The Frequency of the Indel Length Per Strain, Ar: 201, 213, and73

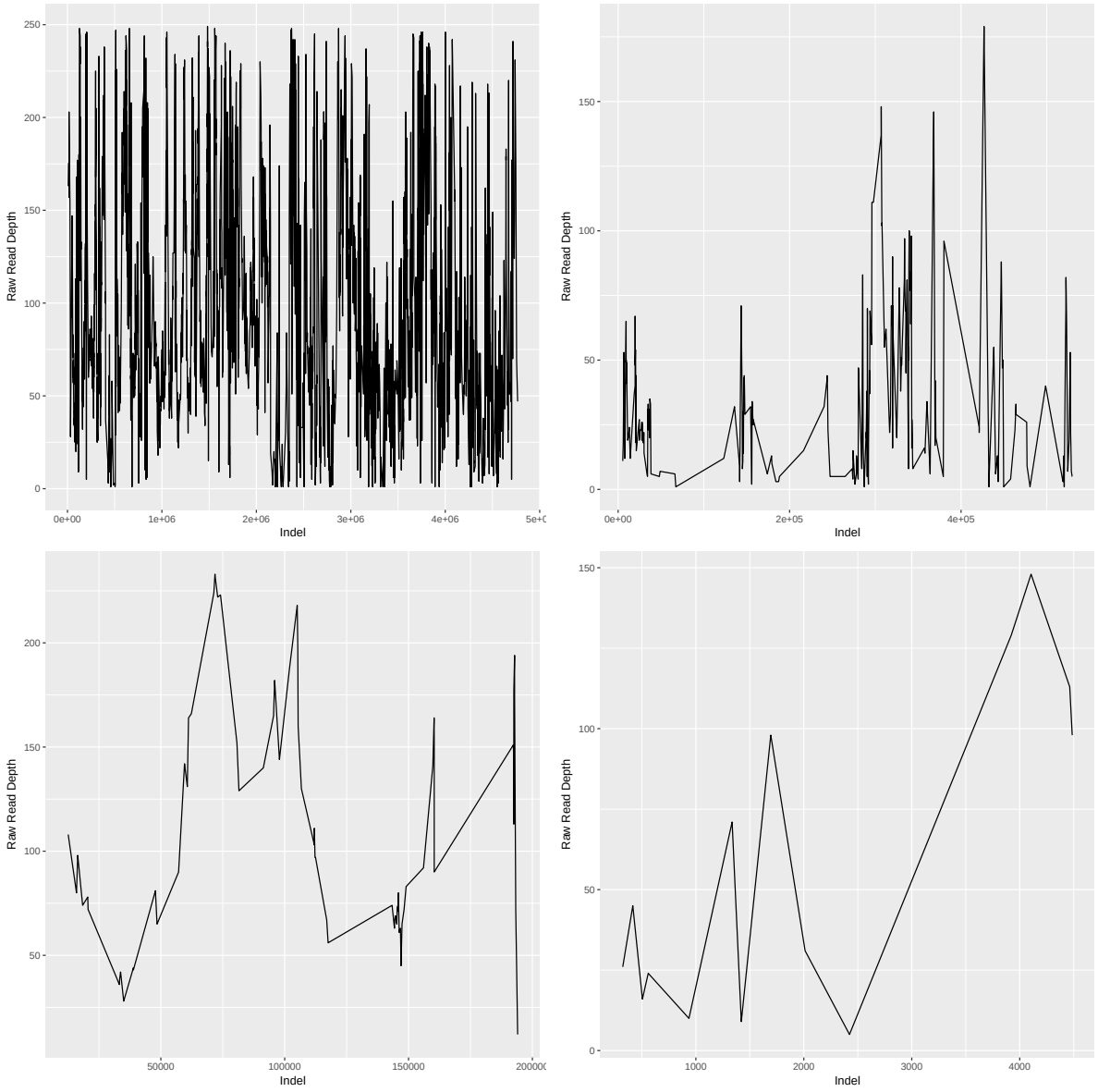

Figure 12. The Read Depth of Indels Found in Ar:109 at Scaffold:1, 50, 100, and 211. Every vertice indicates one indel

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