Armillaria Gallica Gene Analysis Tools

Table of Contents

List of Figures

List of Tables

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`.

Thomas Methods

Read Depth Snapshots

The first work which I completed was the analysis of the read depth values for each of the 15 strains of the *Armillaria gallica*. This work was based on analysis which Hao Wang had done previously to determine locations for each strain that had unusually high read depth. These locations are based on alignment to the 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.

Kassandra Methods

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.

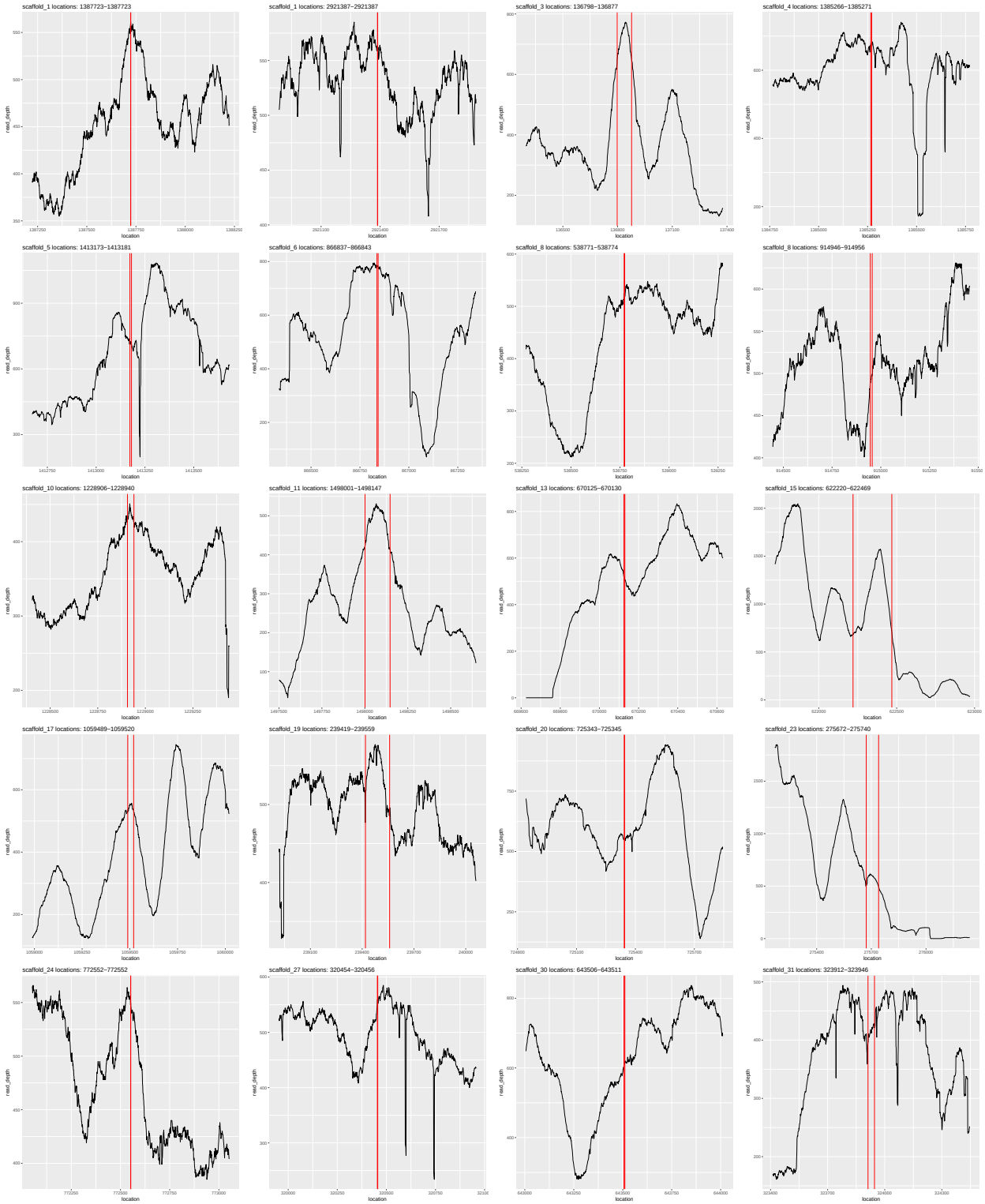

Figure 1. Average Read Depth Snapshots. Graphs are of locations 1, and every 100 starting at 100 until 1900.

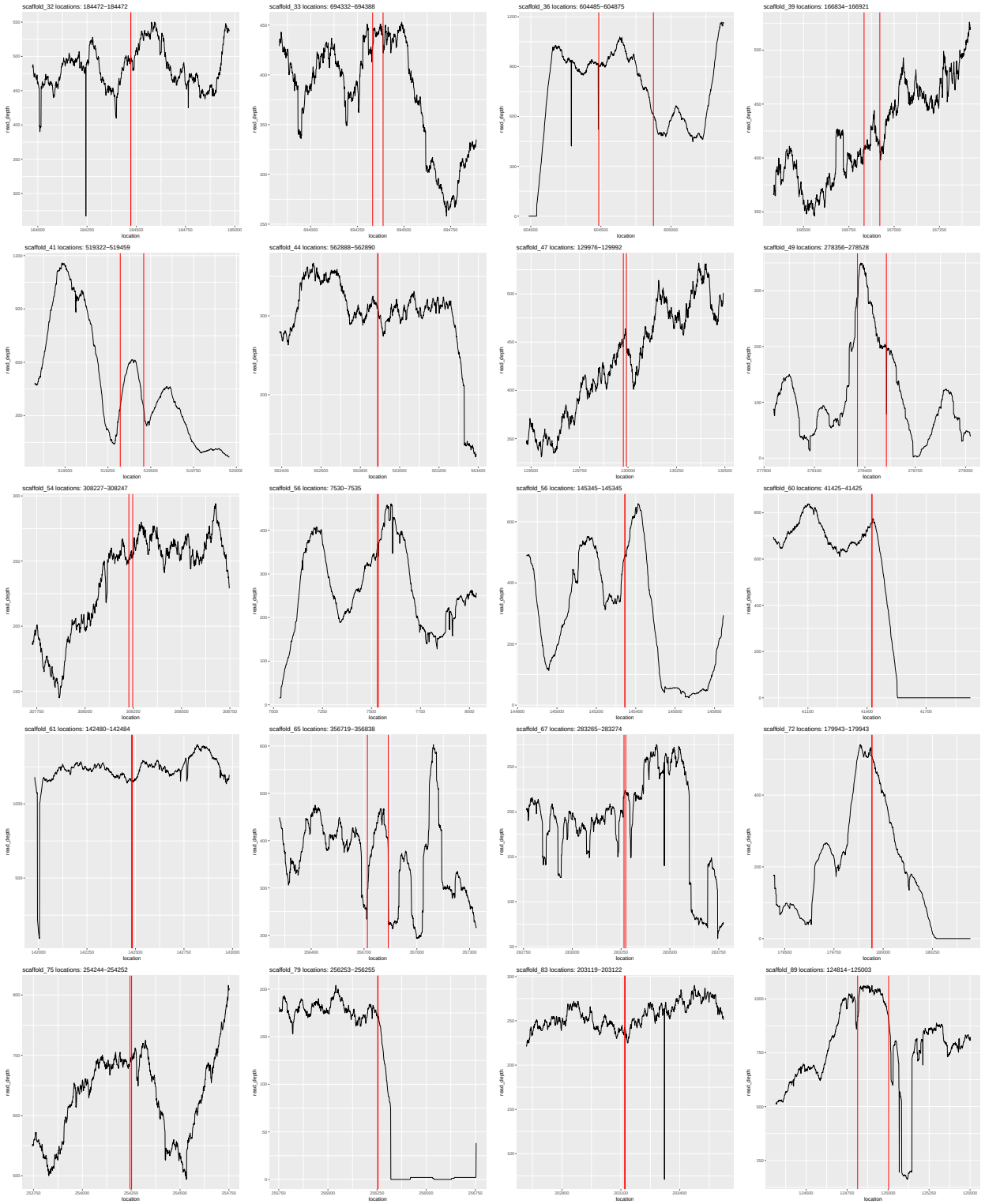

Figure 2. Average Read Depth Snapshots. Graphs are of locations 1, and every 100 starting at 2000 until 3900.

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

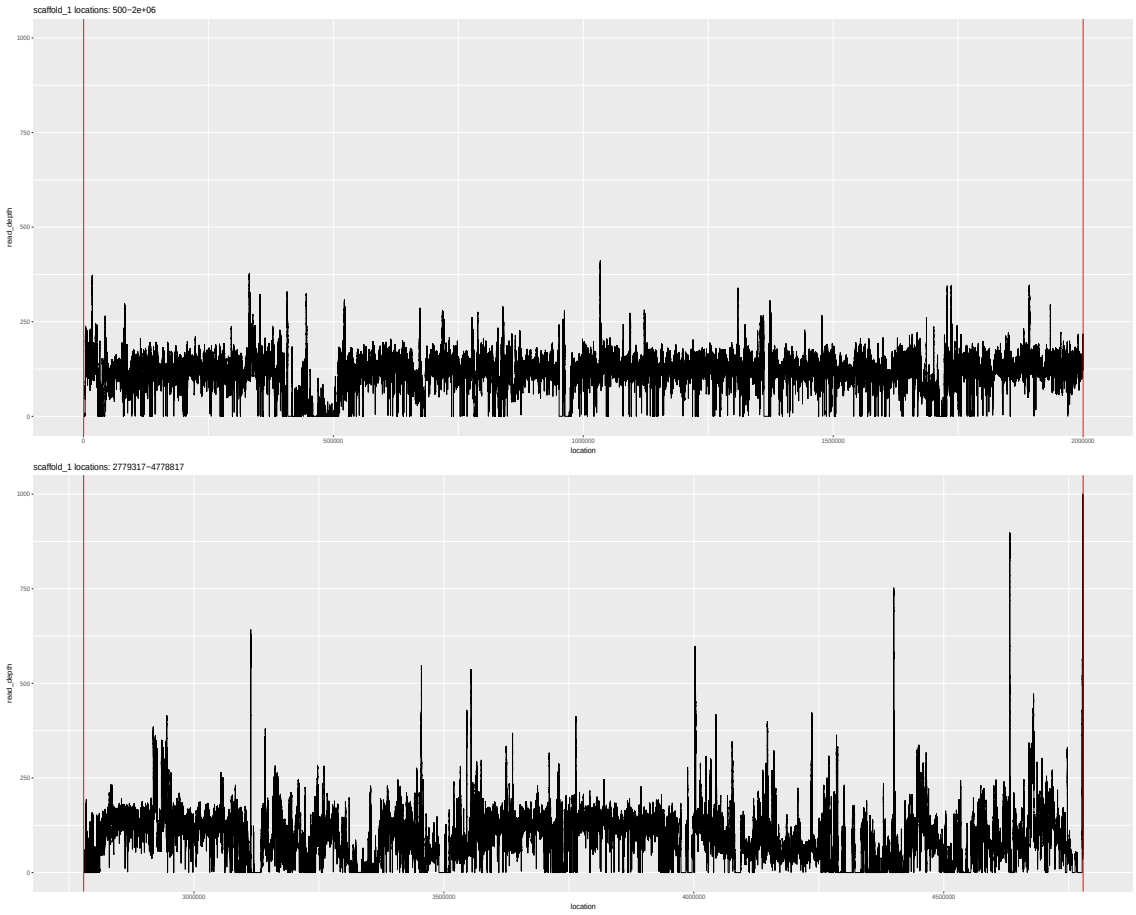

Figure 3. Global Scaffold Read Depth Example for Ar73, scaffold 1.

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 1.

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

Table 1 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.

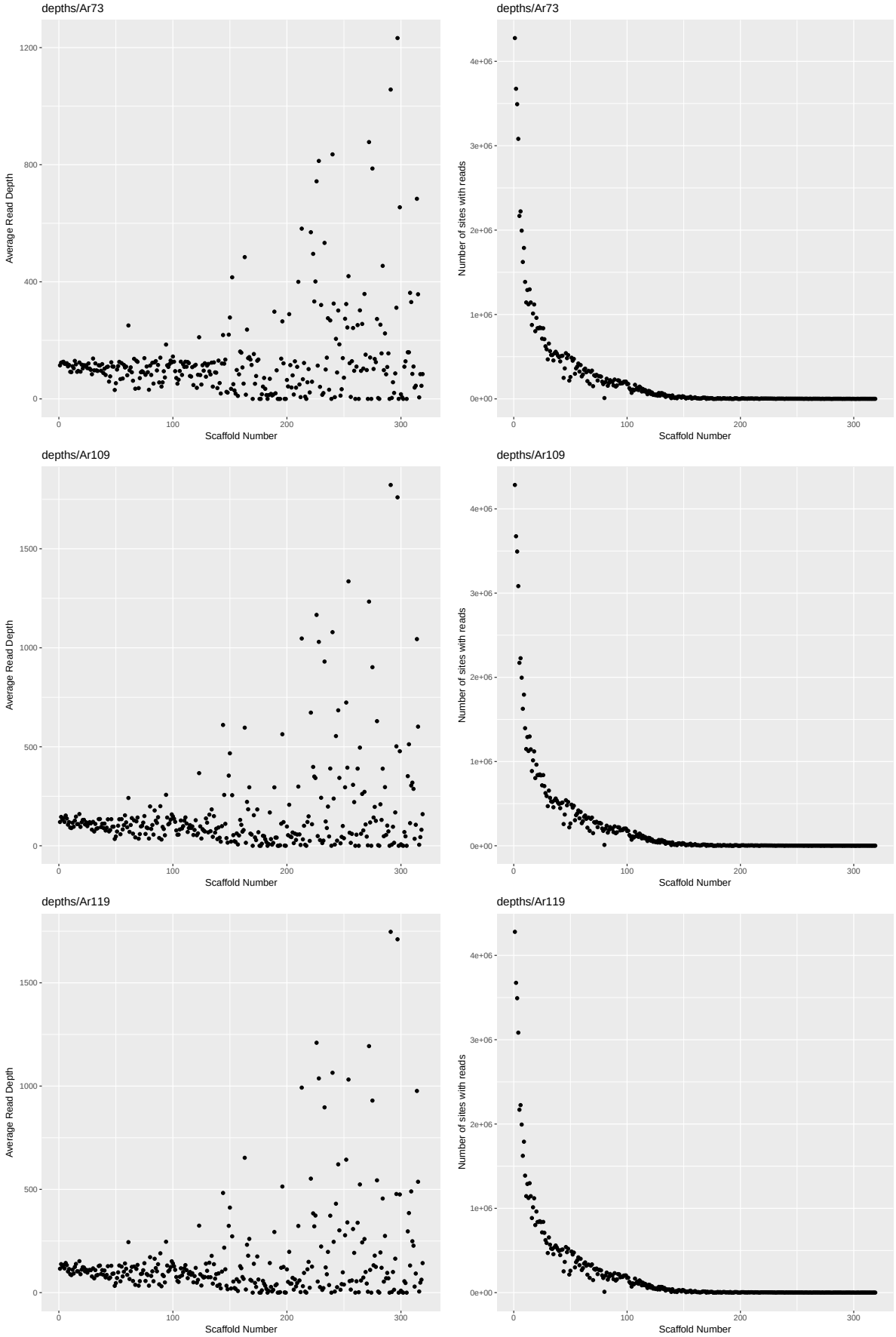

Figure 4. Average Read Depth Per scaffold, Ar:73, 109, 119. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

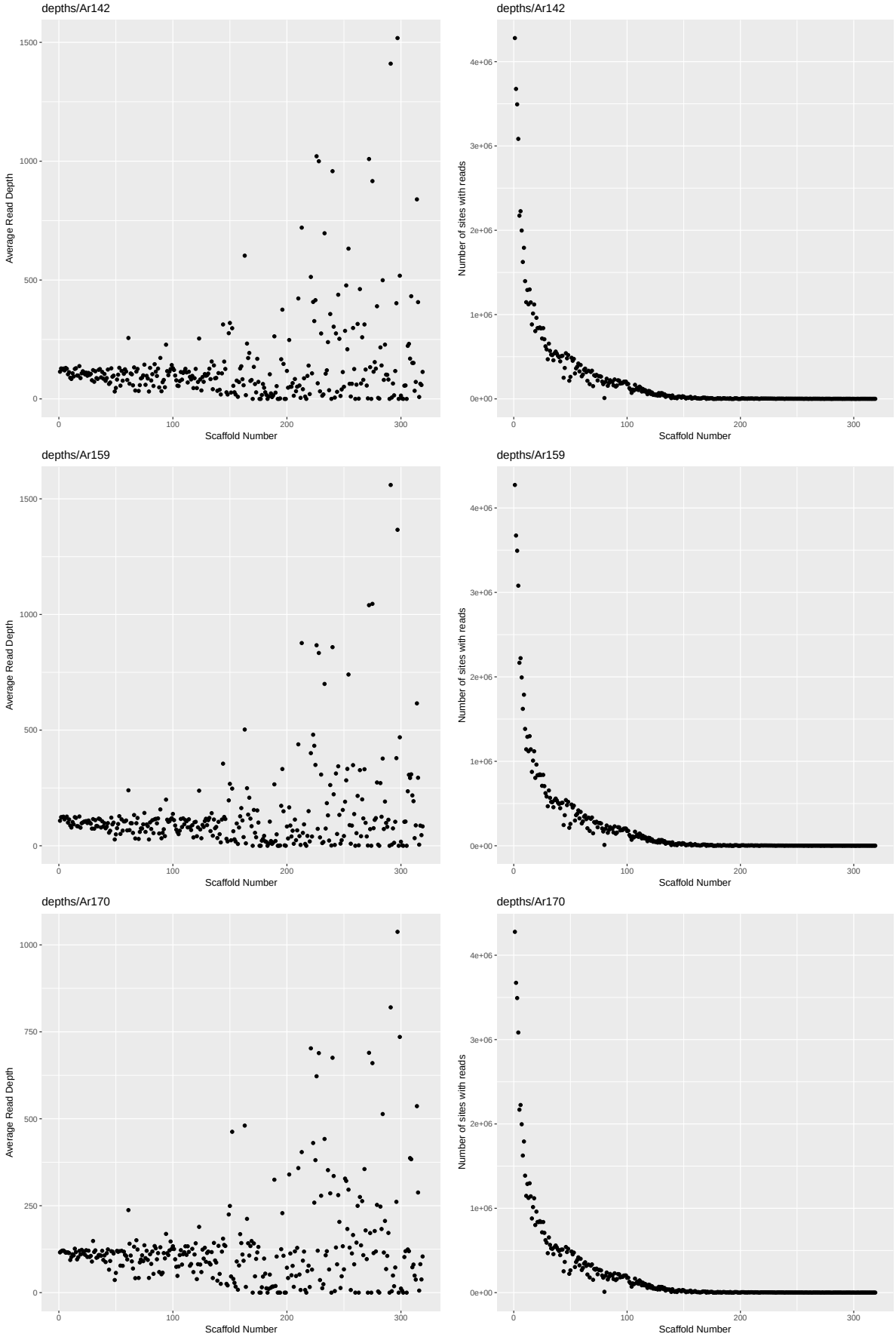

Figure 5. Average Read Depth Per scaffold, Ar: 142, 159, 170. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

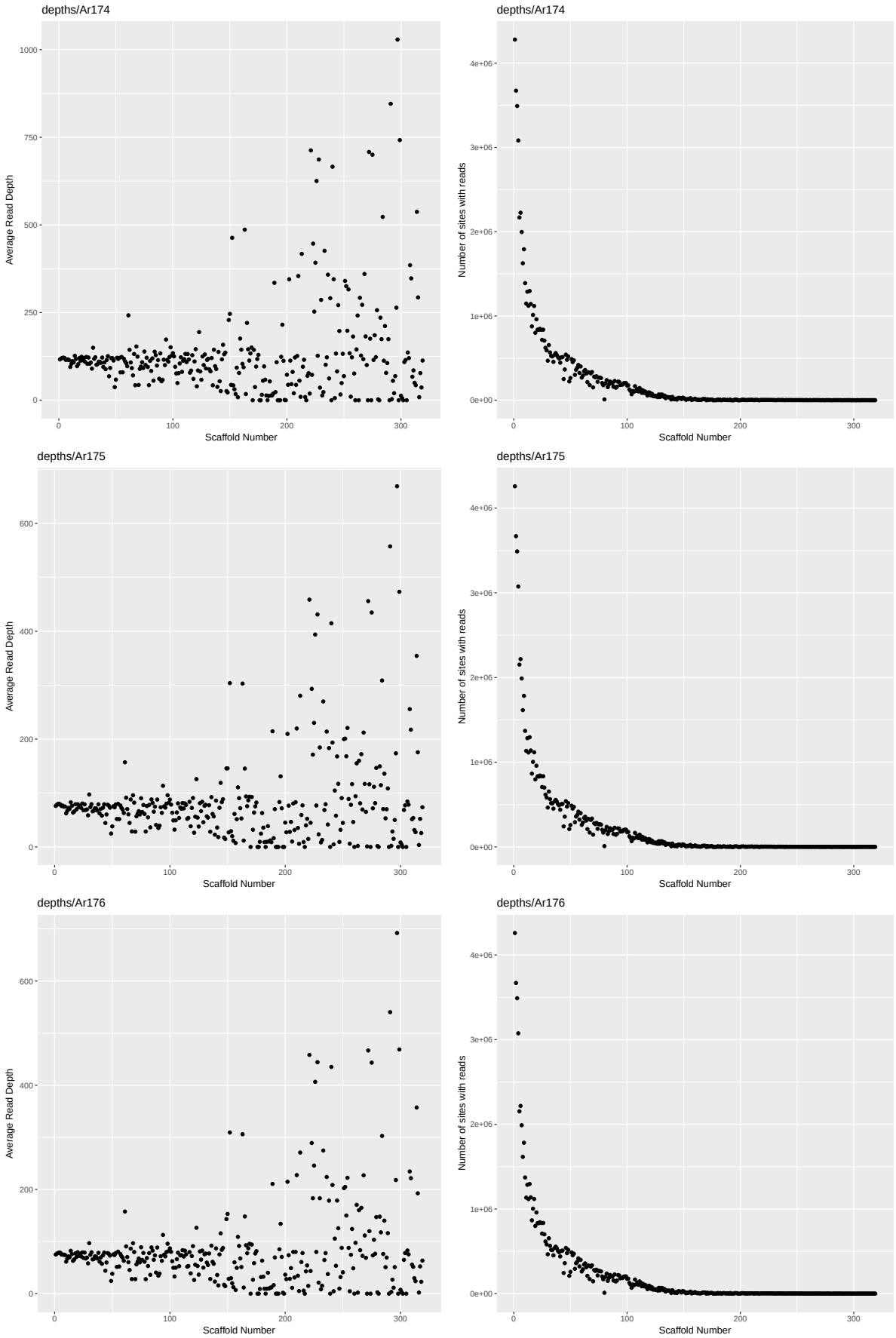

Figure 6. Average Read Depth Per scaffold, Ar:174, 175, 176. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

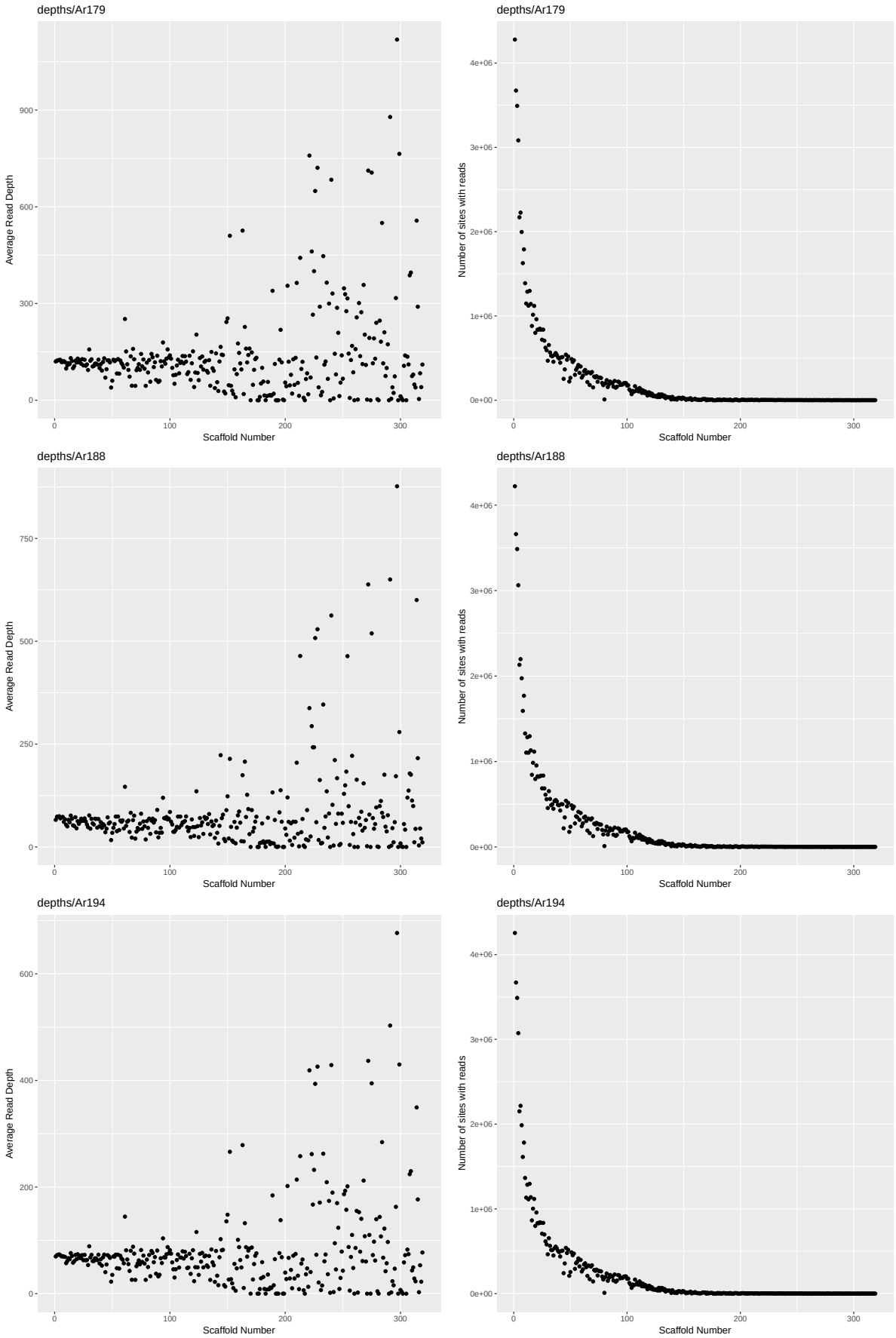

Figure 7. Average Read Depth Per scaffold, Ar:179, 188, 194. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

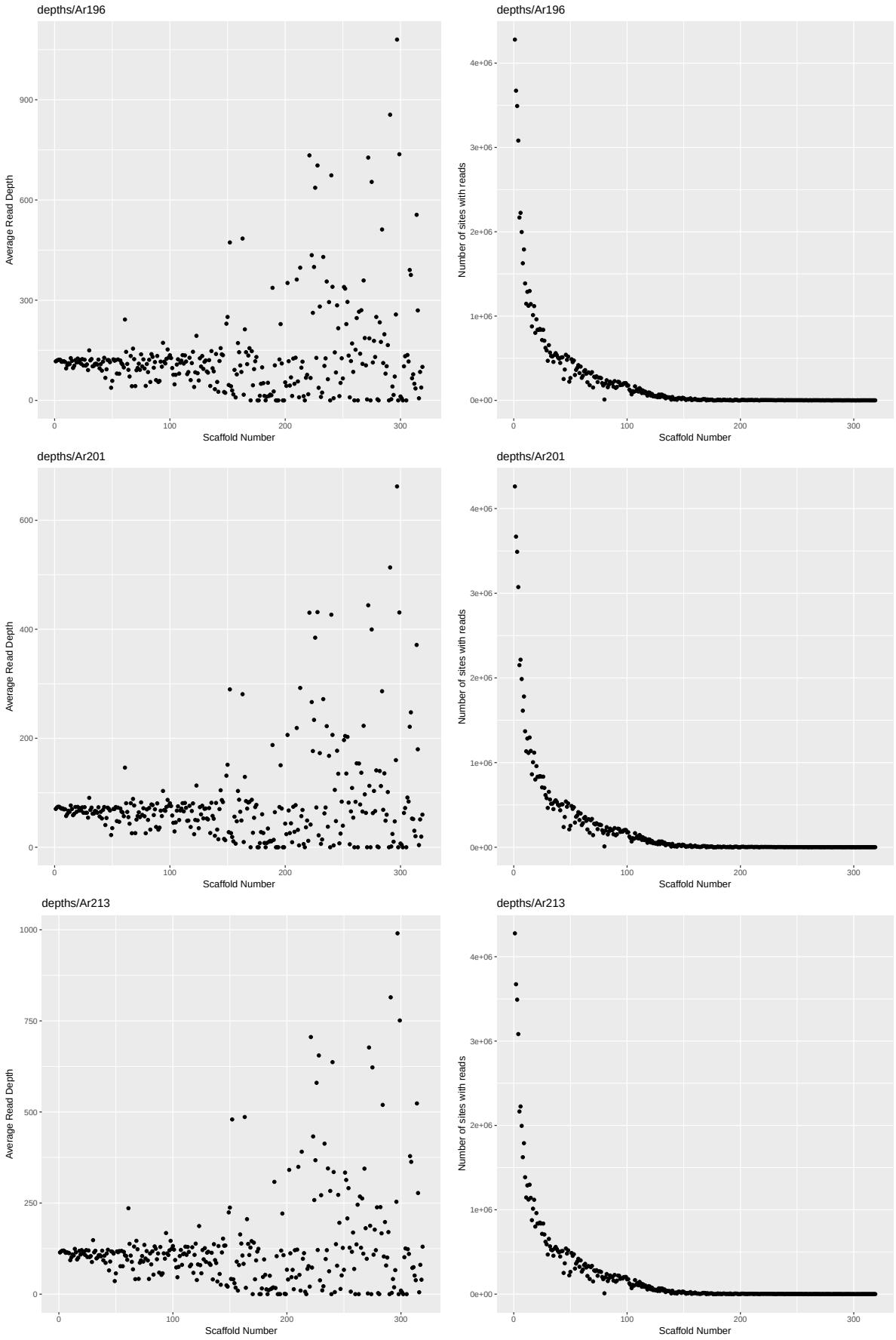

Figure 8. Average Read Depth Per scaffold, Ar:196, 201, 213. Graphs on the left are the average read depth vs scaffold number and graphs on the right are the total number of sites with reads aligned to them per scaffold.

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 using samtools and an additional package called bcftools.

Through the use of bcftools, labels were added to the output which indicated various other information about the indel found such as: maximum number of reads supporting an indel, raw read depth, the number of reads that support the indel and other information.

As seen in 2, 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 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

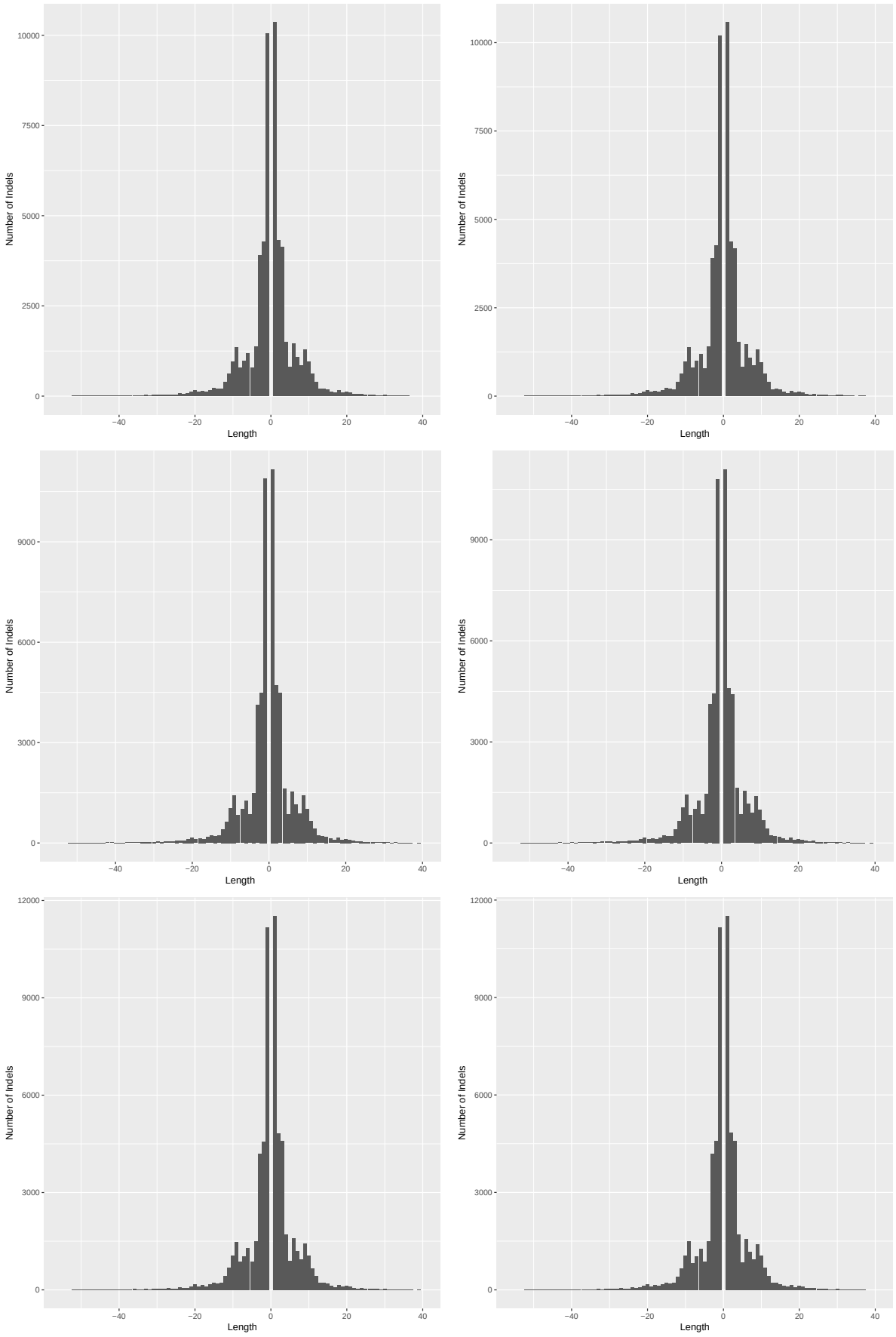

Figure 9. The Frequency of the Indel Length Per Strain, Ar: 109, 119, 142, 159, 170, and 174

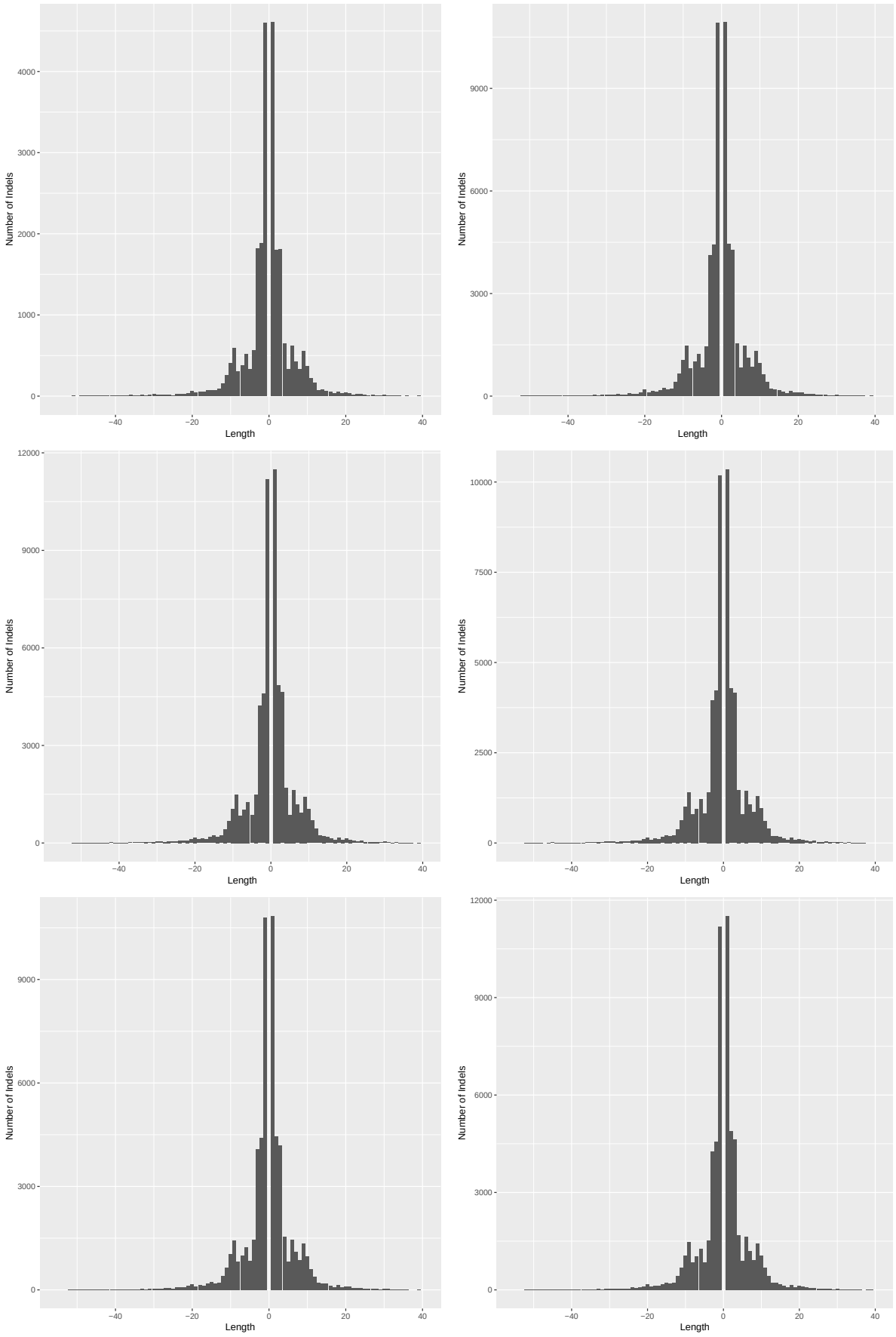

Figure 10. The Frequency of the Indel Length Per Strain, Ar: 175, 176, 179, 188, 194, and 196

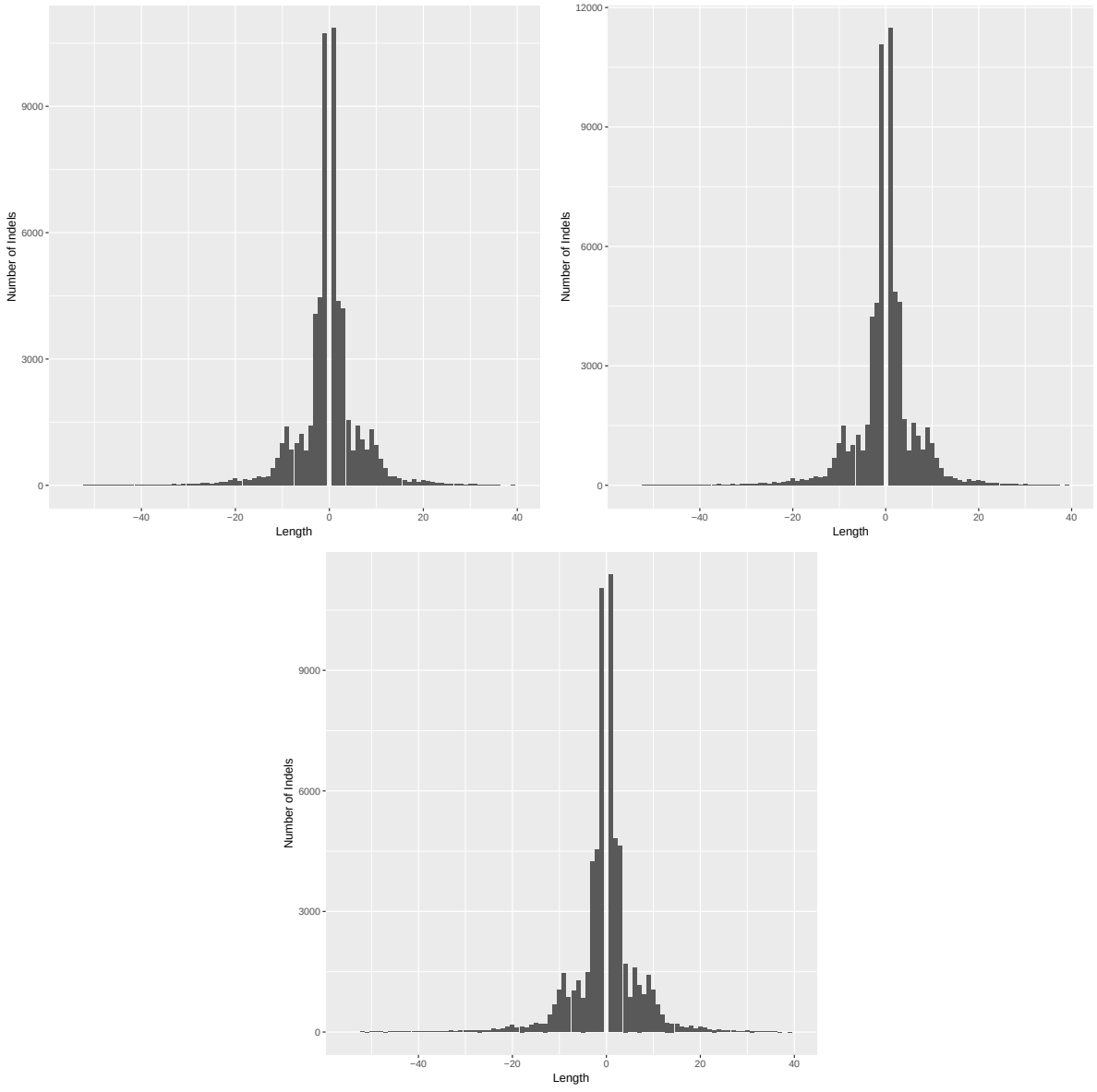

Figure 11. The Frequency of the Indel Length Per Strain, Ar: 201, 213, and73

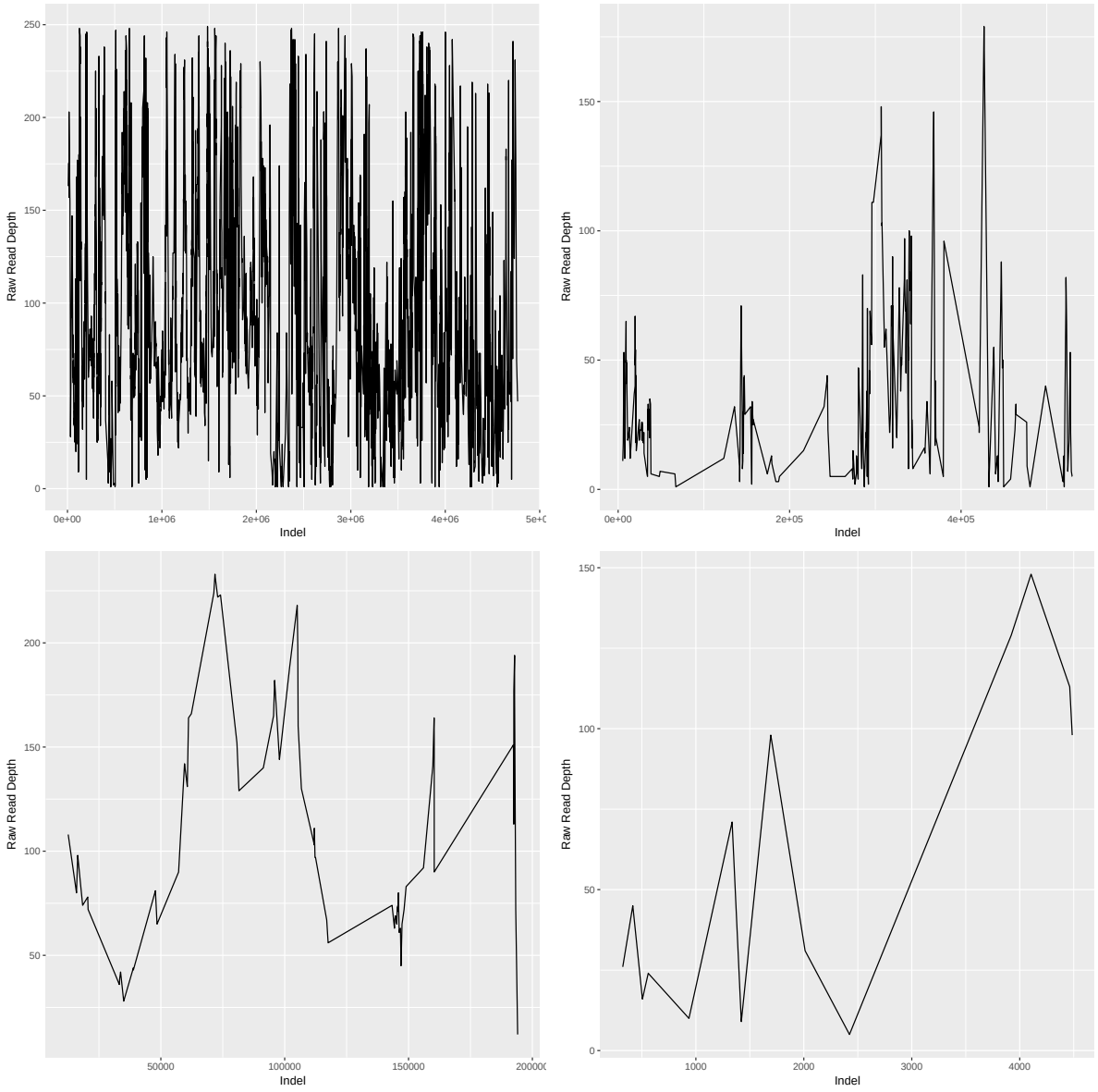

Figure 12. The Read Depth of Indels Found in Ar:109 at Scaffold:1, 50, 100, and 211. Every vertice indicates one indel