Final Mapping Report

Jacob Westaway

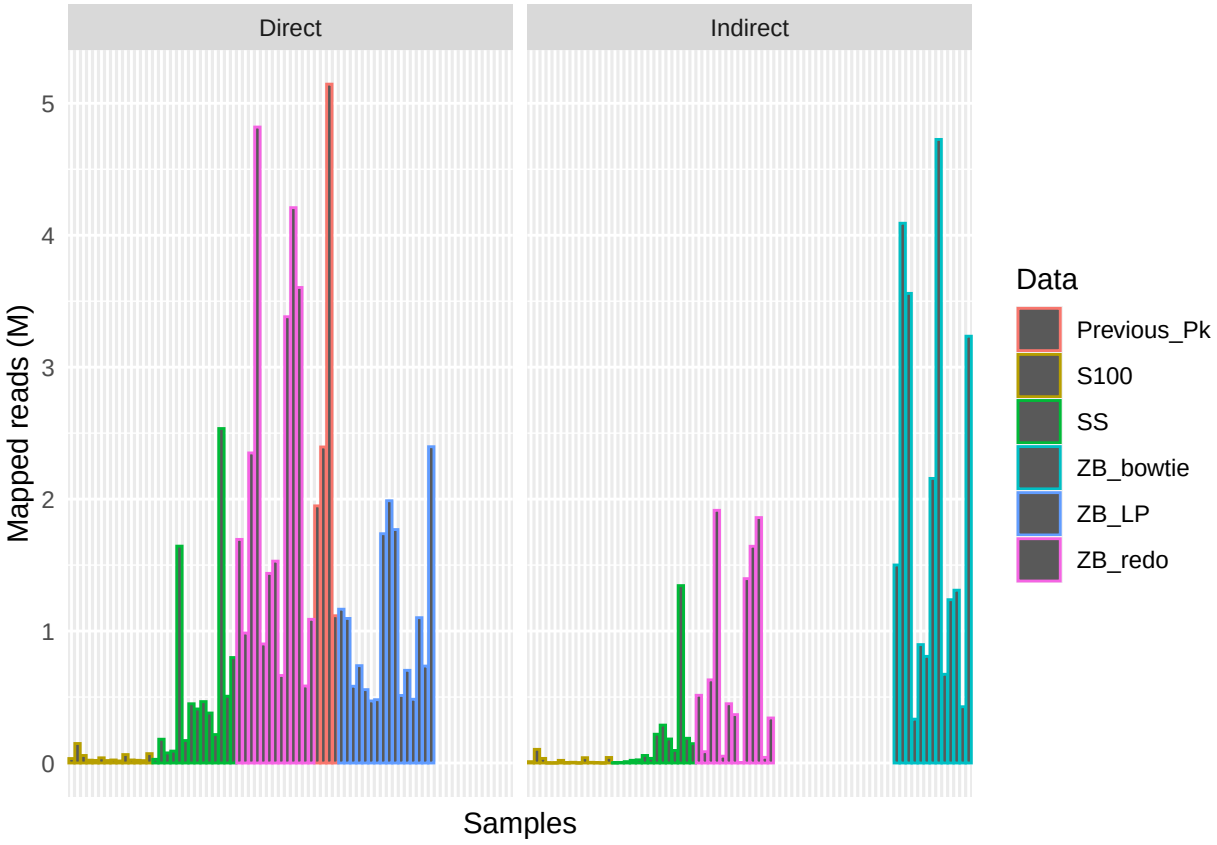
Last updated on 2021-08-02

About

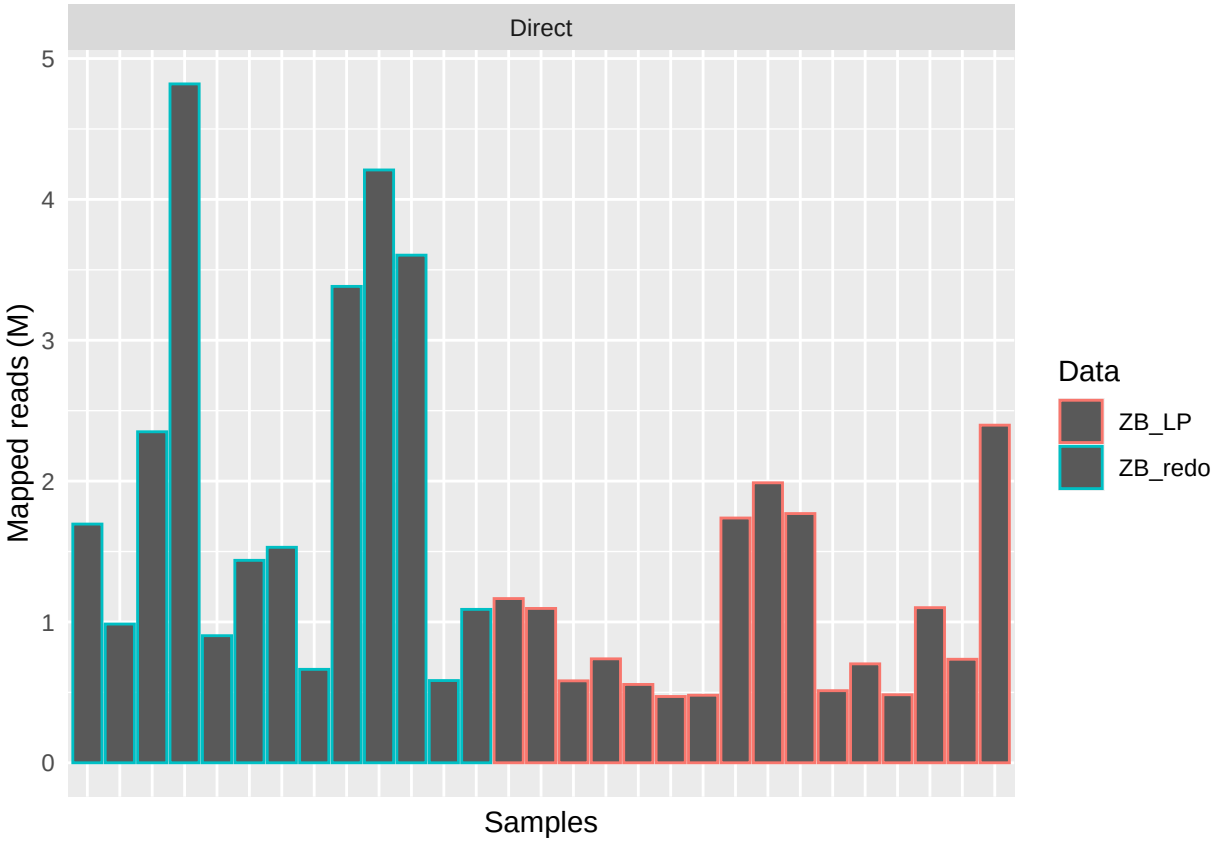
After sharing the previous comparisons with Matt G, it was noted that none of the samples in the previous subset had low parasitemia (< 5000). Thus, another subset of only low parasitemia samples (16) was created, to determine if ZB's methods capture enough depth and coverage at lower parasitemia. Key terms/abbreviations:

- ZB_LP: data from Singapore/Zbynek Bozdech with low parsitemia.
- ZB: data from Singapore/Zbynek Bozdech (truncated).
- ZB_redo: data from Singapore/Zbynek Bozdech (not truncated).
- SS: initial subset from Matt Grigg.
- S100: data from Sanger.
- Previous_Pk: data from a previous study provided by Ernest.
- Direct: aligning/mapping to Pk genome without removal of human contamination.
- Indirect: aligning/mapping to Pk after removal of human contamination.

Comparison of the number reads mapping to Pk across all datasets and for different alignments



Plot a comparisons of mapped reads on ZB data for the indirect alignments with bowtie2 (for removing human contamination) and direct with bwa



Calculate the average mapped reads for both low and 'high' parasitemia

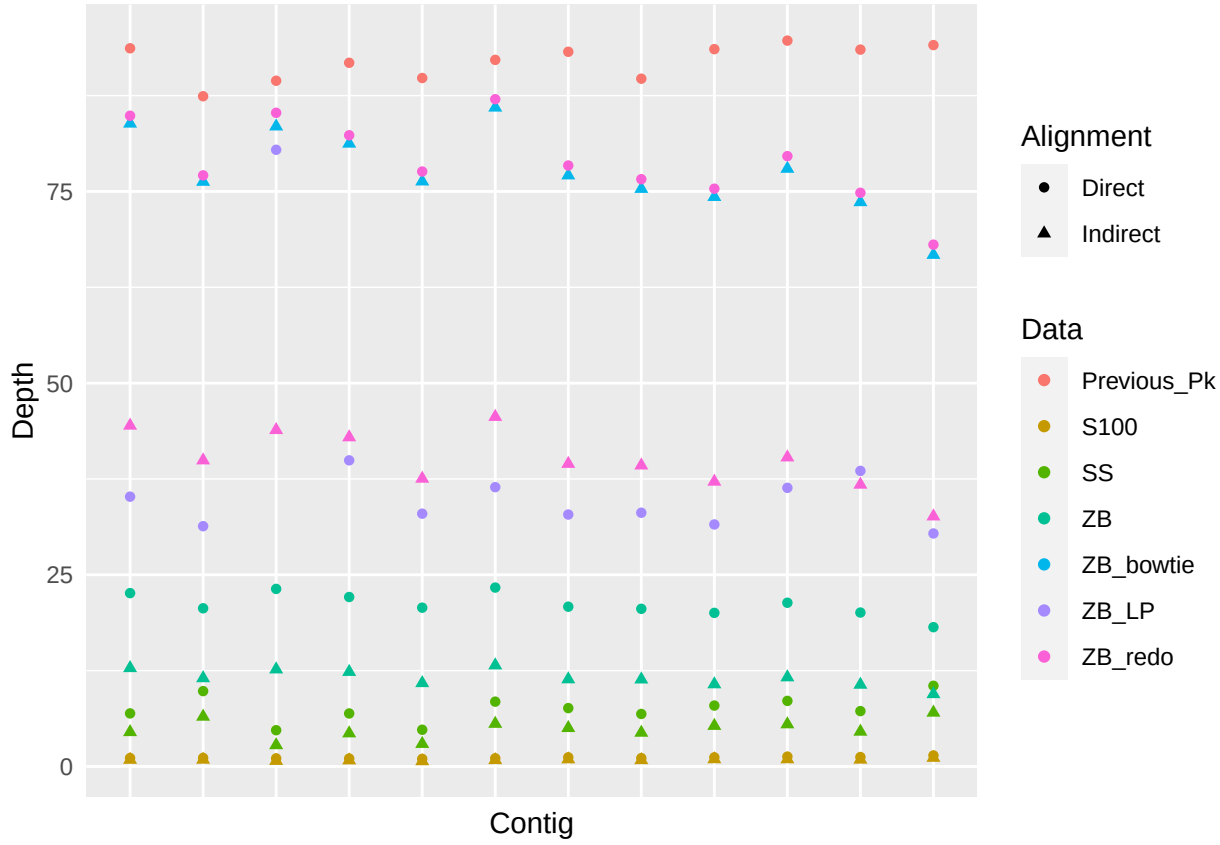
	Low	High
Reads	62302157	64588467
Mapped.reads	10321439	20962735
Mapped.bases	871021885	2020620497
Average.coverage	35.81437	83.08323
Average.coverage.with.deletions	35.81612	83.10338
Percent.of.reference.bases.covered	69.49062	88.19154

Key points:

- On average, the low parasitemia samples have roughly half the number of mapped reads.
- Low parasitemia samples also have less coverage.

Read depth: number of reads aligning at each base

Plot the average read depth per contig



Calculate the average read depth for both low and 'high' parasitemia per contig

Contig	Low	High
PKNH_01	35.19099	84.87679
PKNH_02	31.34139	77.09896
PKNH_03	80.43332	85.25808
PKNH_04	39.92727	82.32885
PKNH_05	32.97133	77.59960
PKNH_06	36.42751	87.03086
PKNH_07	32.86036	78.38100
PKNH_08	33.09138	76.59704
PKNH_09	31.56480	75.35841
PKNH_10	36.34152	79.60490
PKNH_11	38.55282	74.82865
PKNH_12	30.39217	68.06255
PKNH_13	29.43413	68.92501
PKNH_14	28.78352	67.04745

Key points:

- Low parasitemia samples have far lower read depth.

IGV

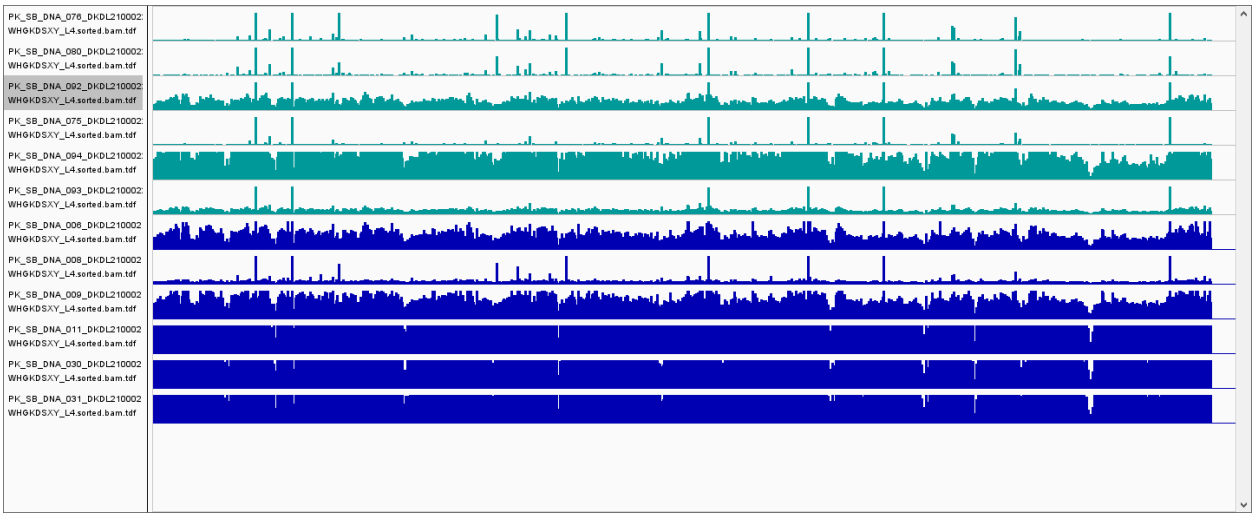
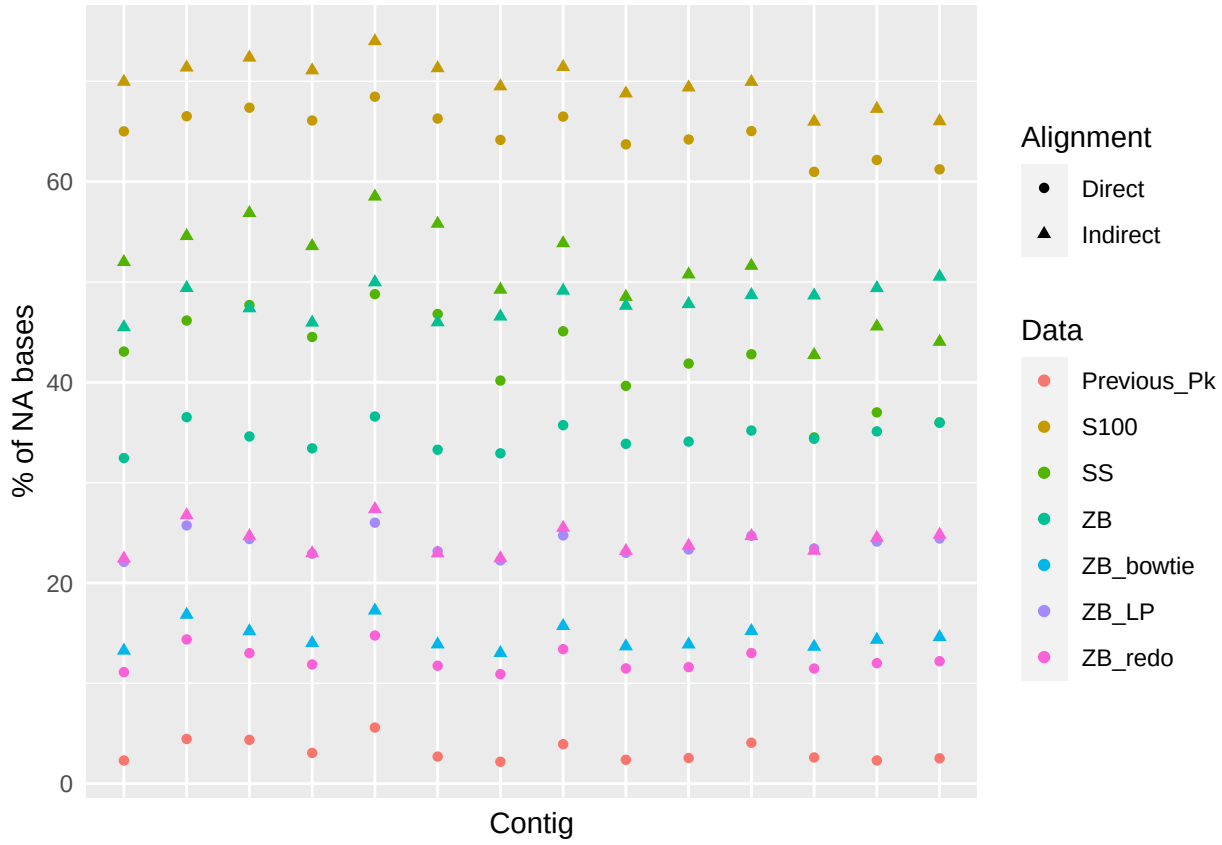


Figure 1: Comparison of read depth across the genome for high and low parasitemia data (direct alignments)

Coloured by data:

- Low parasitemia = light blue.
- High parasitemia = dark blue.

Plot the percentage of bases WITHOUT coverage



Calculate the average percentage of bases without coverage for low and 'high' parasitemia

Contig	Low	High
PKNH_01	22.10345	11.11487
PKNH_02	25.73536	14.37436
PKNH_03	24.38272	13.00764
PKNH_04	22.91200	11.88241
PKNH_05	26.01125	14.76420
PKNH_06	23.16447	11.75098
PKNH_07	22.24840	10.91841
PKNH_08	24.74369	13.40286
PKNH_09	23.01070	11.48517
PKNH_10	23.34298	11.61447
PKNH_11	24.69642	13.01593
PKNH_12	23.43585	11.47563
PKNH_13	24.12781	12.00943
PKNH_14	24.43757	12.20479

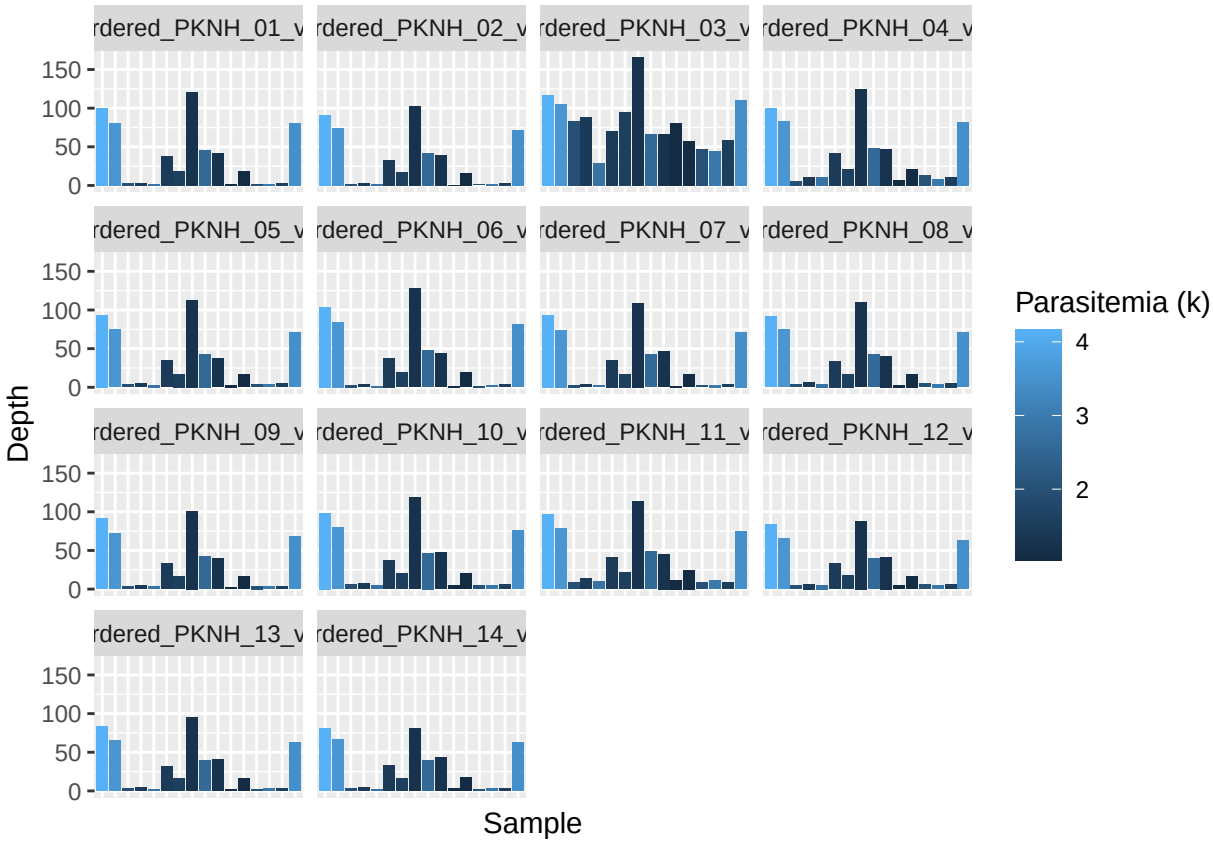
Key points:

- Low parasitemia samples map to less of the genome.

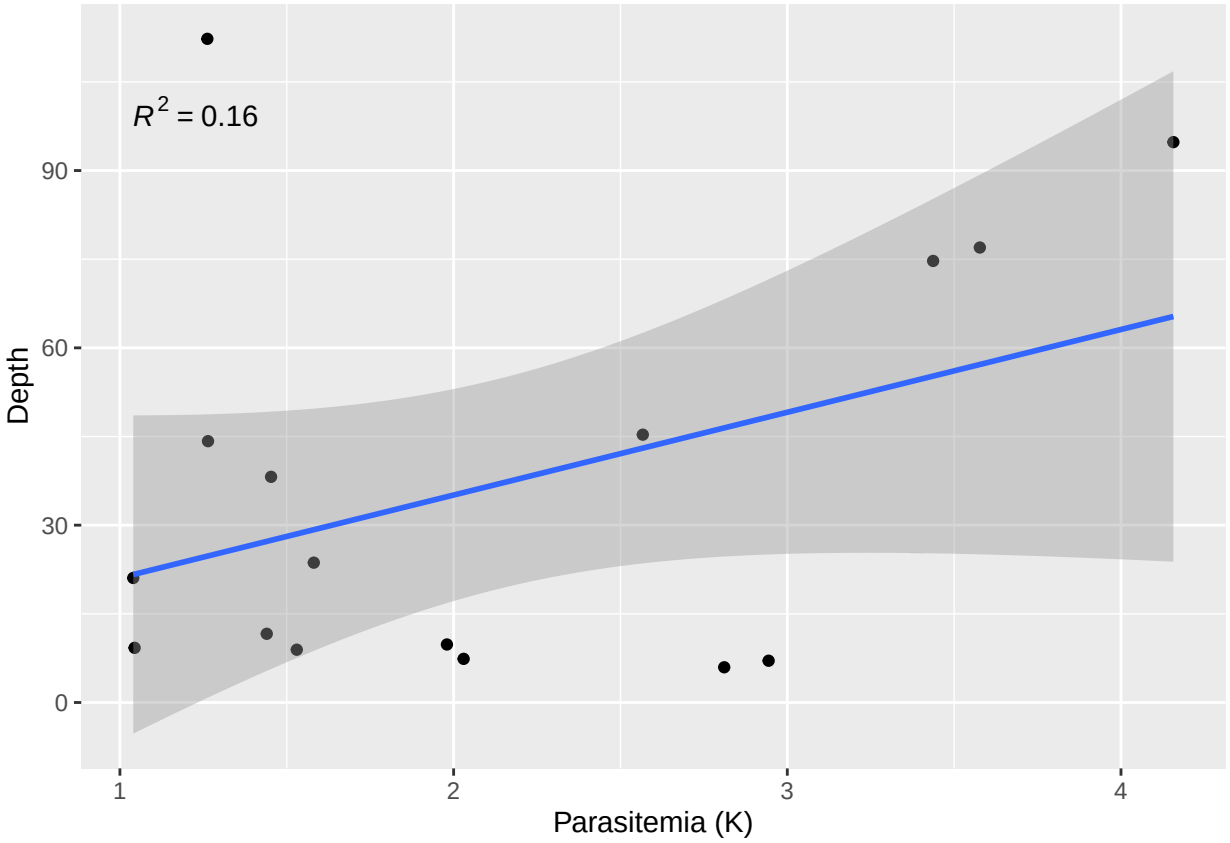
Effect of parasitemia on read depth in low parasitemia samples

Read in metadata and combine with summary data to explore the effect of parasitemia on read depth

Explore the relationship between parasitemia read depth across samples



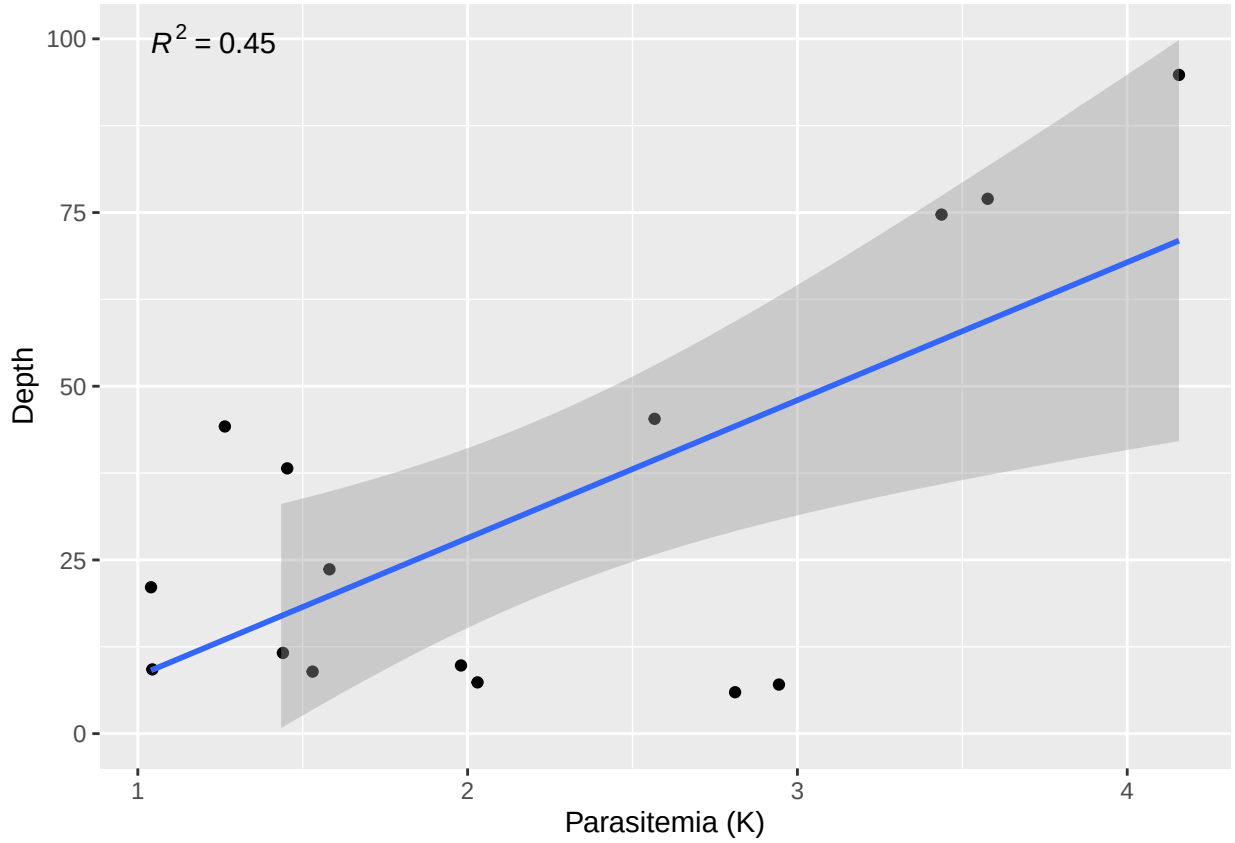
Plot relationship between depth across samples in relation to parasitemia



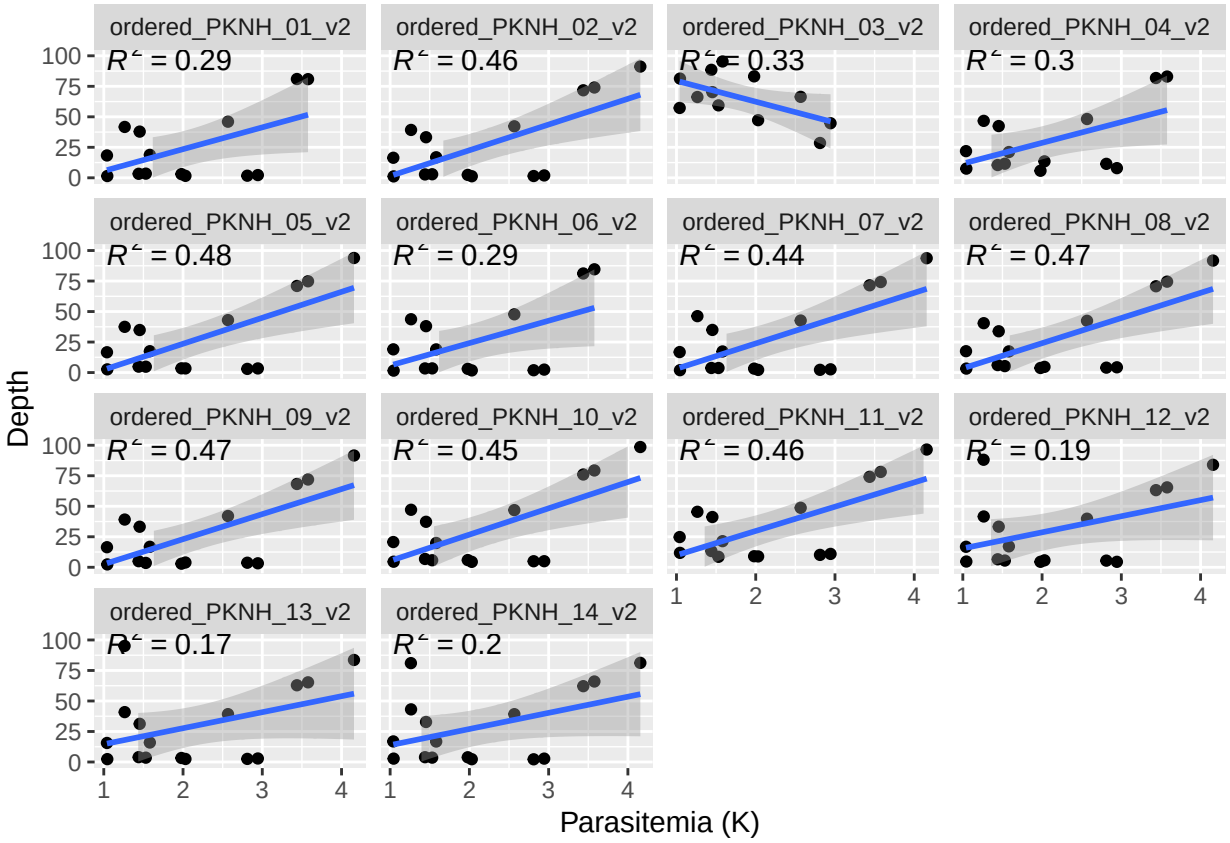
Key points:

- The outlier (ID = 70) in the last two plots is interesting, having the highest read depth and one of the lowest paraistemias of the 'low'paraistemia samples. I checked out several metrics, and the number of duplicates, all other quality metrics, reads and mapped reads are all comparable to other samples. So I'm not sure why we are getting such good depth for this low parasitemia sample.

Plot relationship between depth across samples in relation to parasitemia - outlier removed



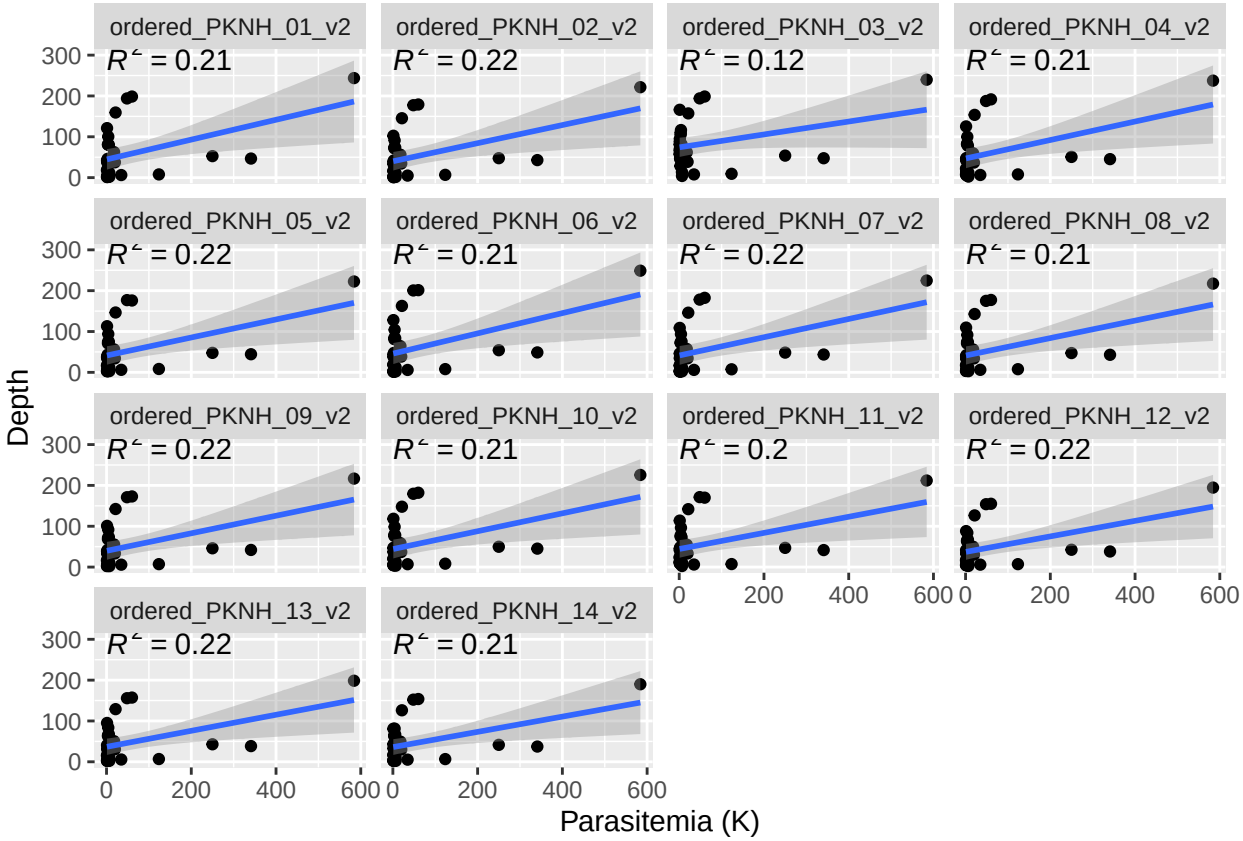
Plot relationship between depth across samples in relation to parasitemia, faceted by contig and outlier removed



Key points:

- Moderate relationship between read depth and parasitemia for low parasitemia samples.

Compare read depth and parasitemia across both low and high parasitemia



Summary of mapping statisitcs, ordered by parasitemia

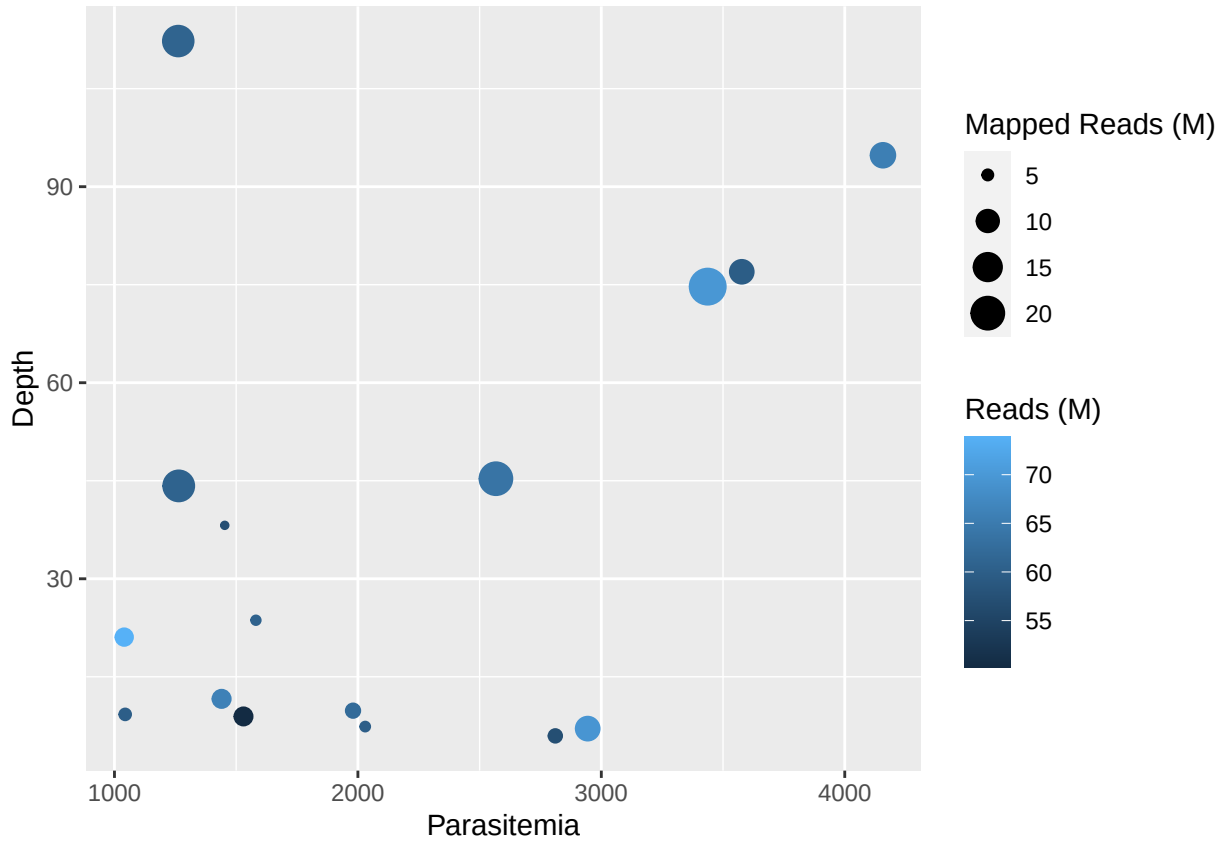
ID	Reads	Mapped.reads	Percent.mapped	Depth	Parasitemia
076	73858676	7024811	9.511	21.070746	1040.00
075	59958738	5118642	8.537	9.245902	1044.00
070	60914700	17379057	28.530	112.284528	1262.00
074	60758322	17694310	29.122	44.212491	1264.00
063	65971272	7379423	11.186	11.621332	1440.00
065	57116811	4705862	8.239	38.177380	1453.00
093	50249766	7348807	14.625	8.930064	1530.00
066	60582605	4797638	7.919	23.653912	1581.00
061	62202281	5818429	9.354	9.814160	1980.00
080	59974937	4831837	8.056	7.380827	2030.00
073	63718510	19879597	31.199	45.310426	2567.00
064	57174594	5562223	9.728	5.957668	2811.00
092	69368532	11013785	15.877	7.065444	2944.00
094	69643922	23971740	34.420	74.705439	3437.00
058	59641254	10958761	18.374	76.976119	3577.00
057	65699593	11658102	17.745	94.807859	4157.00
009	68641810	23501174	34.237	74.835392	5522.00
008	60161476	9850931	16.374	11.782071	6450.00

ID	Reads	Mapped.reads	Percent.mapped	Depth	Parasitemia
019	64389558	6634489	10.304	2.640097	6607.00
006	63478370	16948578	26.700	56.567076	17834.00
034	55339553	10892058	19.682	34.603467	19759.00
023	72279366	33823852	46.796	144.815271	21368.00
033	67368656	5838670	8.667	6.126351	34778.64
031	47086256	36040788	76.542	176.228271	48339.00
030	70575585	42089050	59.637	178.268707	60000.00
012	68025812	9023684	13.265	7.625737	123590.40
014	65440877	14374321	21.965	47.996346	249994.20
015	66491135	15299426	23.010	43.210637	340953.80
011	70371619	48198532	68.491	220.941714	584014.90

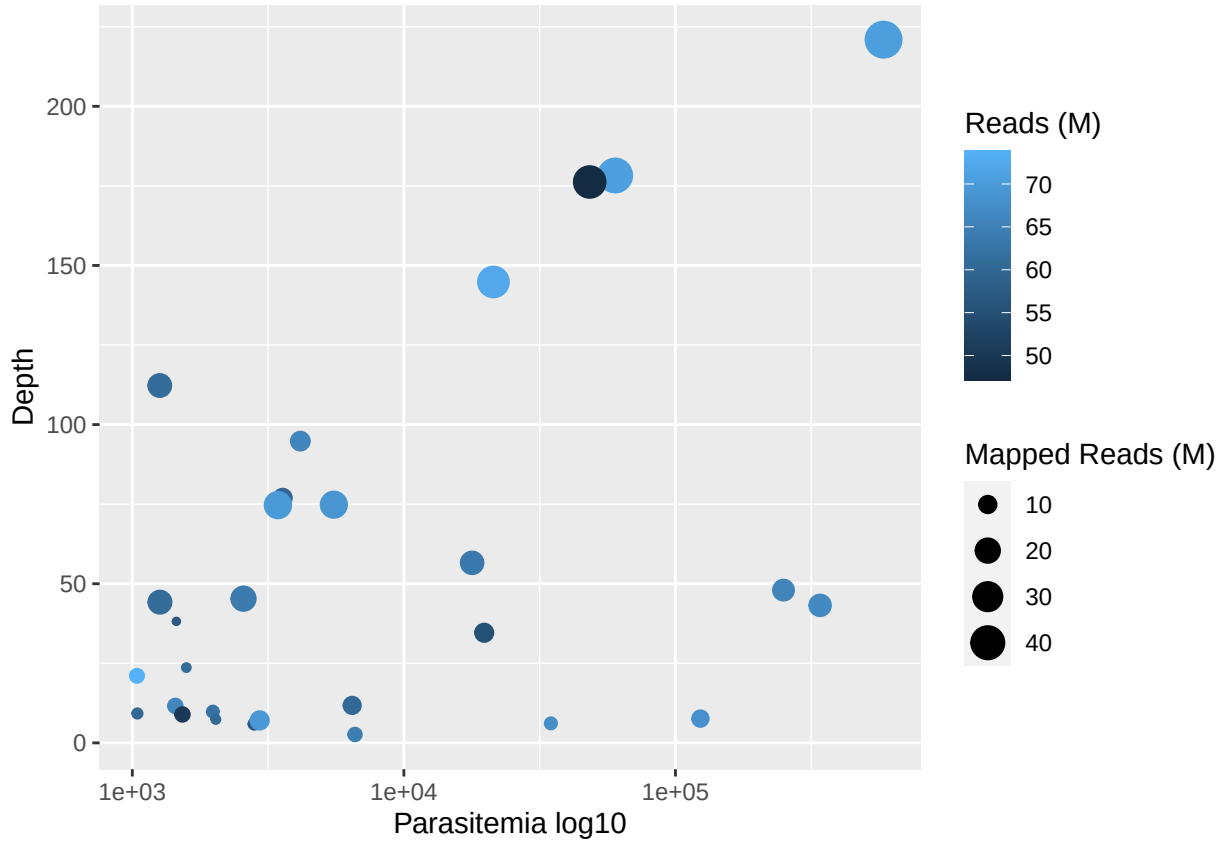
Avergaes for high vs low parasitemia

Data	High	Low
Reads	64588467	62302157
Mapped.reads	20962735	10321439
Percent.mapped	32.74385	16.40138
Average.coverage	83.08323	35.81437
Percent.bases.covered	88.19154	69.49062
Parasitemia	116862.380	2132.312
Depth	77.35701	36.95089

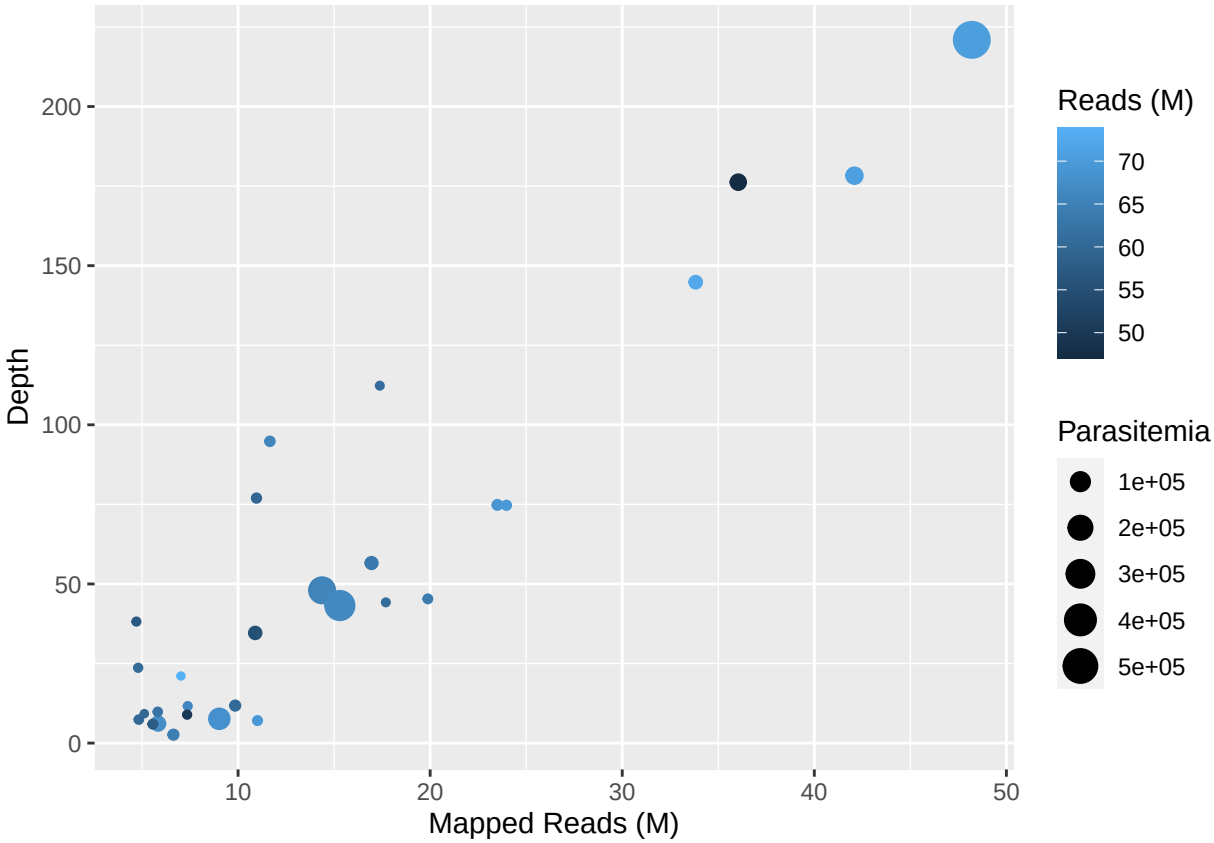
Plot the relationship between reads, mapped reads, read depth and parasitemia in the low parasitemia samples



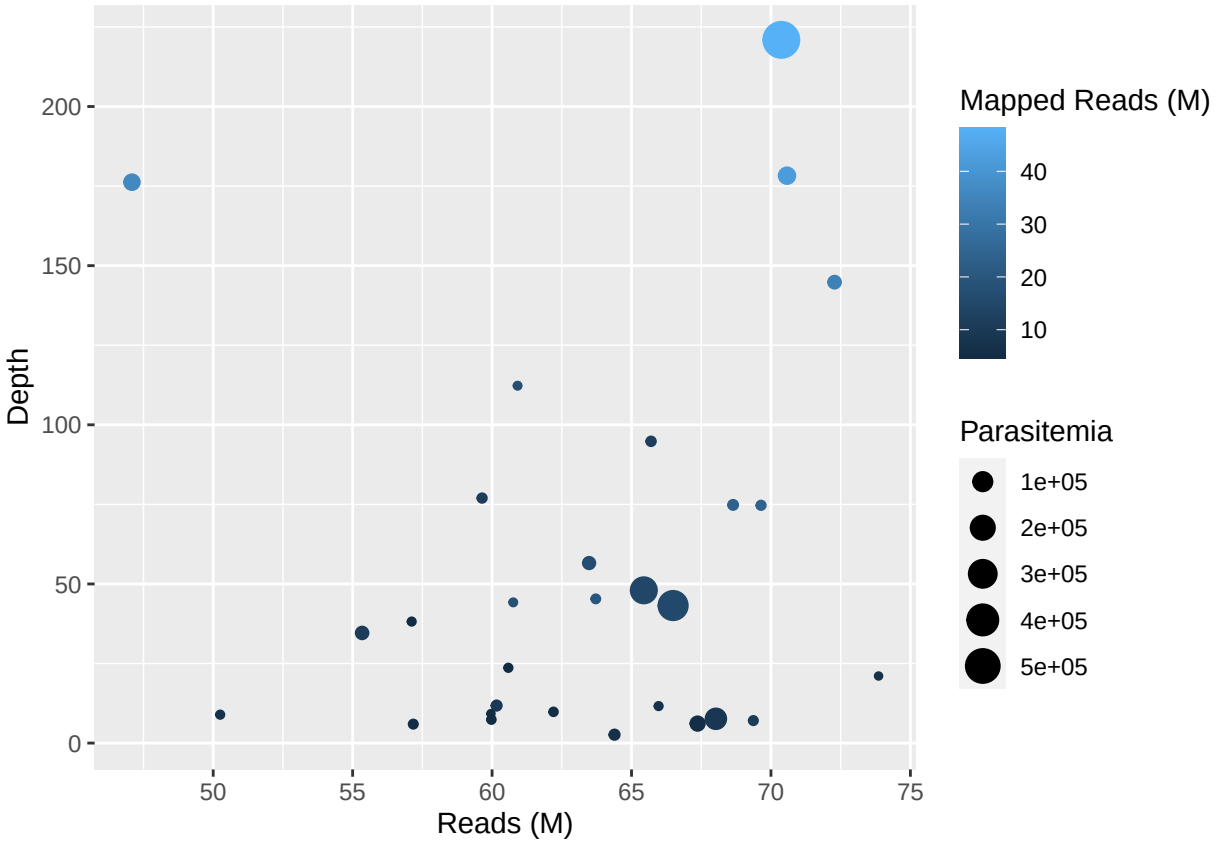
For the entire dataset



Depth vs mapped reads



Depth vs reads



Key points:

- Another outlier on the far left, with low reads and yet high depth (sample 31). Has a relatively high parasitemia of 48339.

Conclusions:

- The read depth is far lower in low parasitemia samples, as there appears to be a low to moderate correlation between parasitemia and read depth.
- The average read depth of ~30 in the low parasitemia samples should be enough to call variants.
- The lower read depth in some low parasitemia samples, those that are maybe below ~2132.312 (mean for low parasitemia), may result in their exclusion from anlayses.