Mapping Report

Jacob Westaway

Last updated on 2021-07-20

About

This report briefly goes through some of the comparisons we have made over the last couple of months to determine what might be the best combination of data (sample preperationa) and alignment type. In brief, we ran three datasets: the initial datset provided by Matt, data from Sanger and data from ZB. We tried aligning directly to the Pk genome (with *bwa*), and indirectly via removal of human contamination (with *bwa*/*bowtie2*). We ran mapping statistics (*bbmap*) and read depth calculations (*samtools*) on this data to get the number of reads aligning and the depth of reads at each position along the Pk genome. We also imported some of this data into *IGV*/*tablet* to do additonal visual comparisons, and downloaded some data from a previous study provided by Ernest to see how our best outputs (ZB's data) compared to a previously successful workflow. Lastly, we used metadata to explore the relationship between ZB's data and parasitemia.

From this work it seems safe to say that **ZB**'s data, along with aligning **directly** to the Pk genome with bwa, is the best approach.

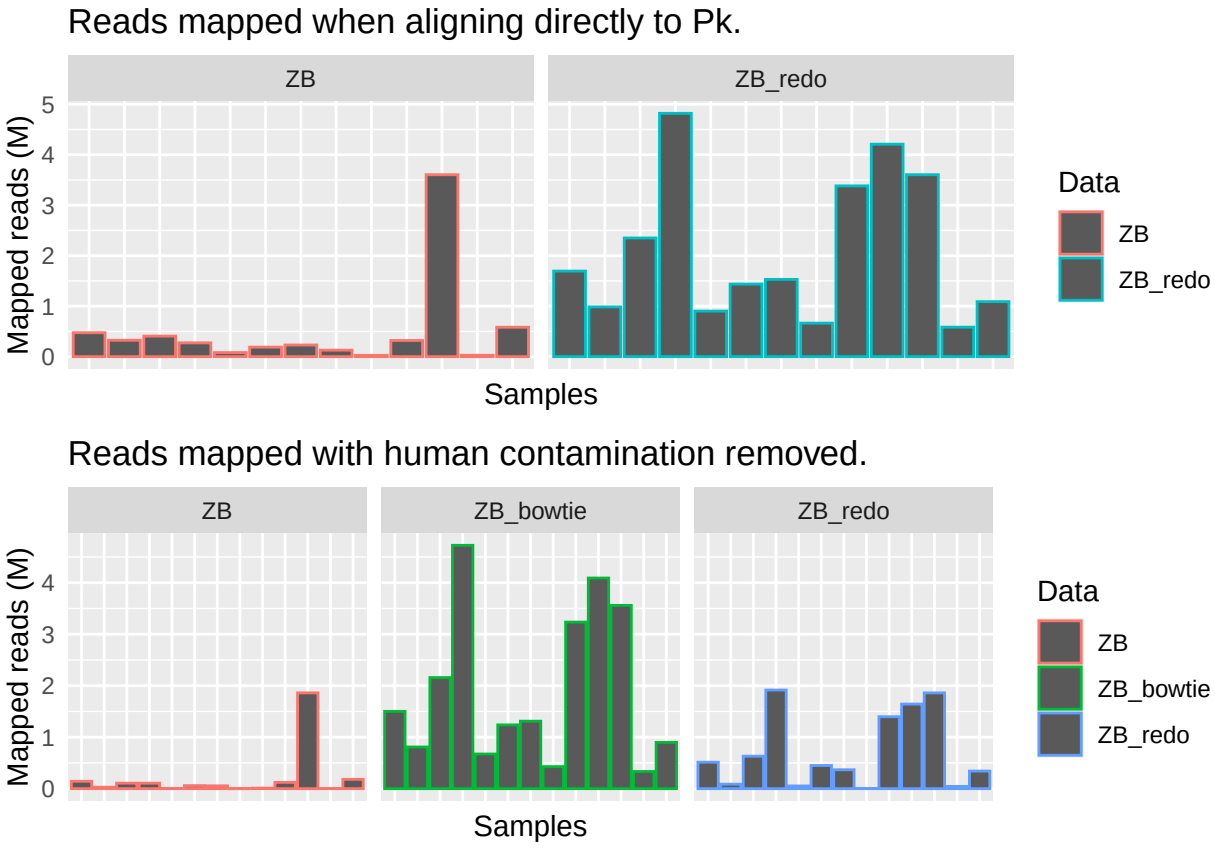
Key terms/abbreviations:

- 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.

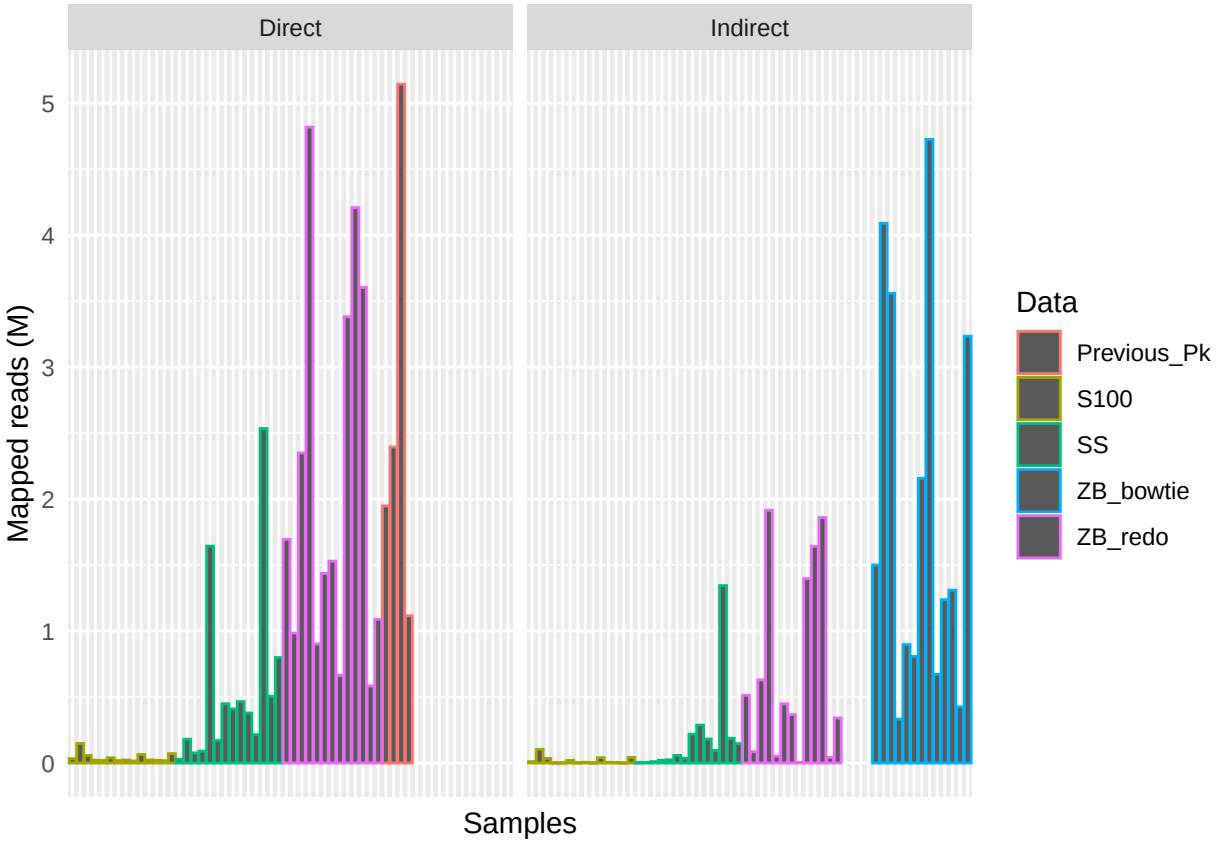
Map Stats: number of reads mapping to reference genome

Comparison across ZB dataset

Firstly, Field and I found an error in my trimming step on the ZB dataset that meant that I was working with truncated files. This resutled in running the ZB data through the workflow again. The plot below shows the difference it made on the number of reads mapping, with ZB_redo and ZB_bowtie using the 'corrected' dataset. Both outputs (ZB and ZB_redo) are included in some subsequent plots for more comparison.



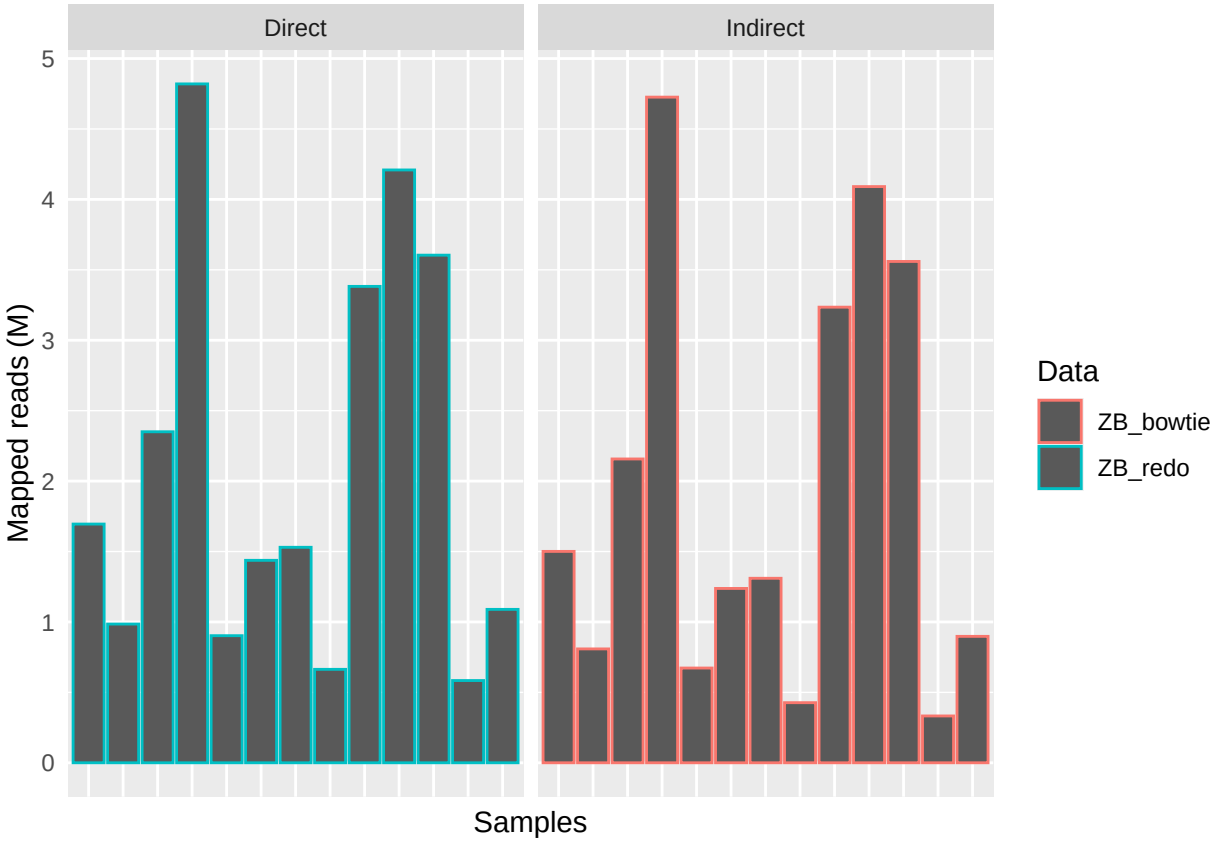
Comparison of the number reads mapping to Pk across all datasets and for different alignments



Key points:

- ZB's data (ZB_redo) appears to get the most reads.
- ZB's data is comparable to 'Previous Pk' data.
- Mapping directly to Pk with bwa & indirectly via removal of the human genome with bowite2 is comparable.

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 these two alignmetns

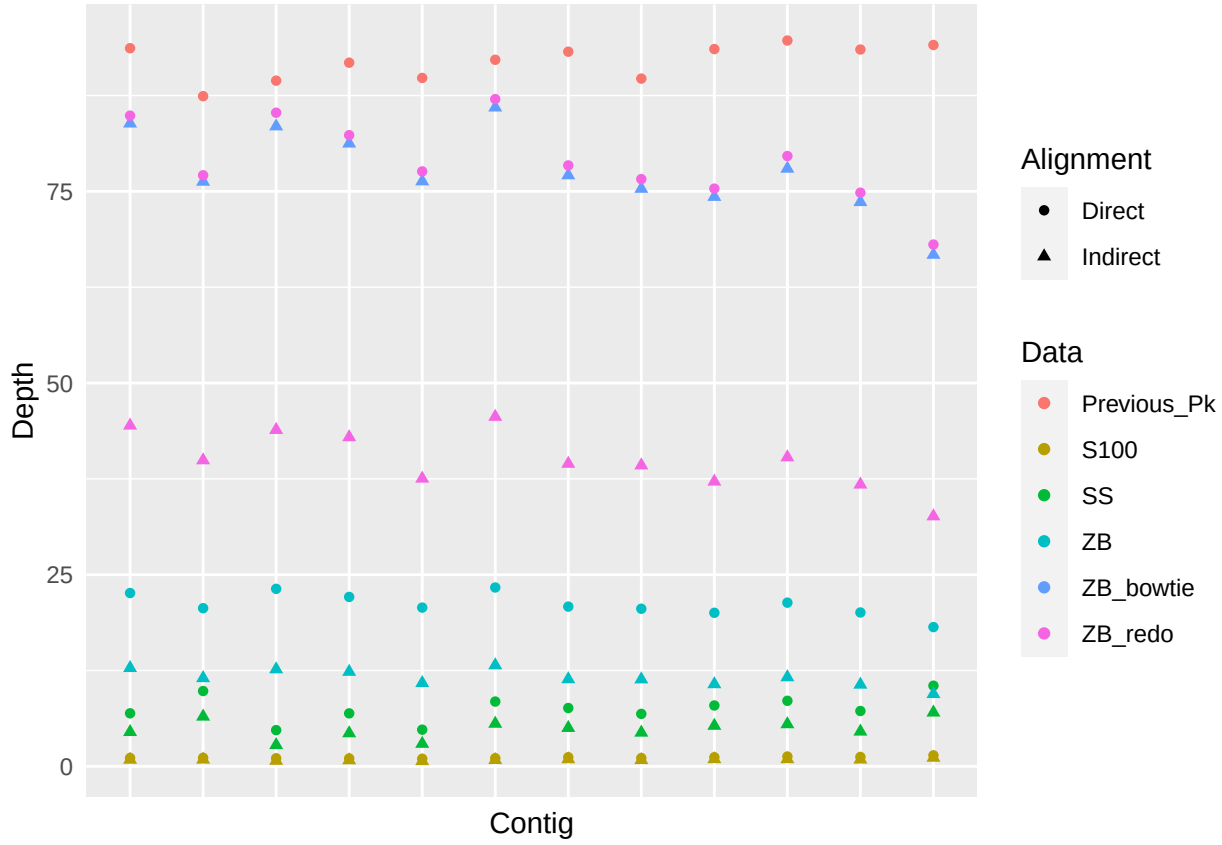
Data	ZB_bowtie	ZB_redo
Reads	27234378	64588467
Mapped.reads	19195718	20962735
Mapped.bases	1978467141	2020620497
Average.coverage	81.35008	83.08323
Average.coverage.with.deletions	81.37100	83.10338
Percent.of.reference.bases.covered	85.99923	88.19154

Key points:

- Comparable, but mapping directly to Pk with bwa leads to more mapped reads and more coverage.

Read depth: number of reads aligning at each base

Plot the average read depth per contig



IGV

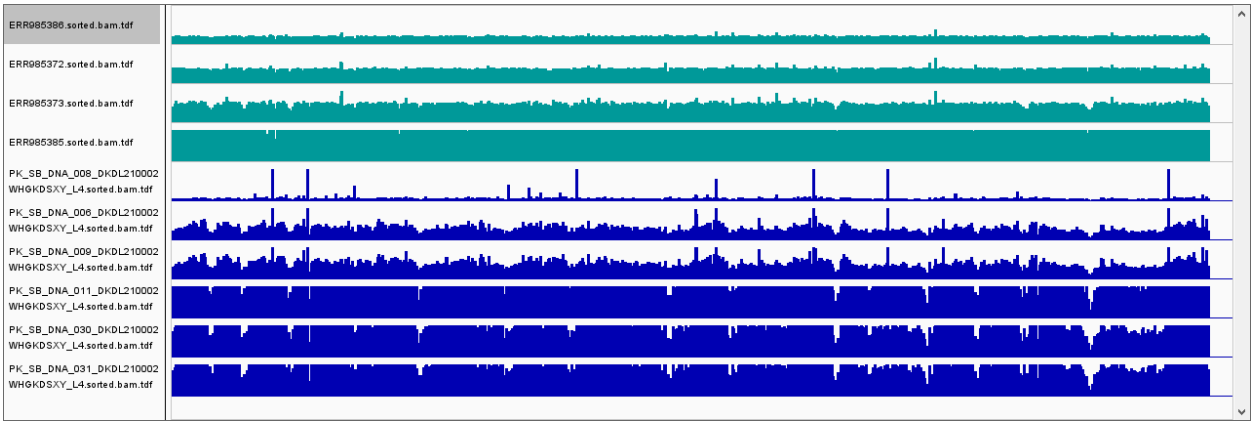


Figure 1: Comparison of read depth across the genome for ZB_redo and Previous_Pk data (direct alignments)

Coloured by data:

- Previous Pk data = light blue.
- ZB data = dark blue.

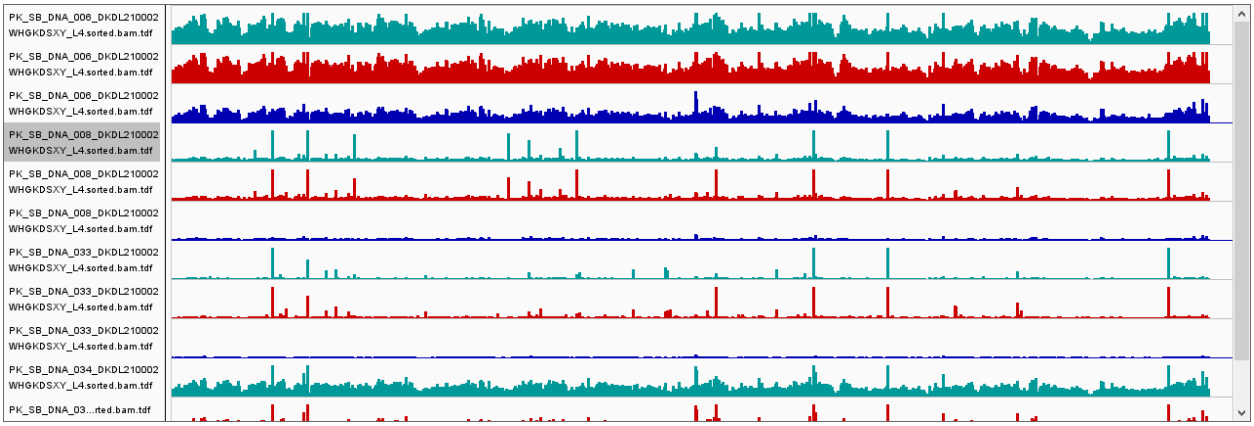
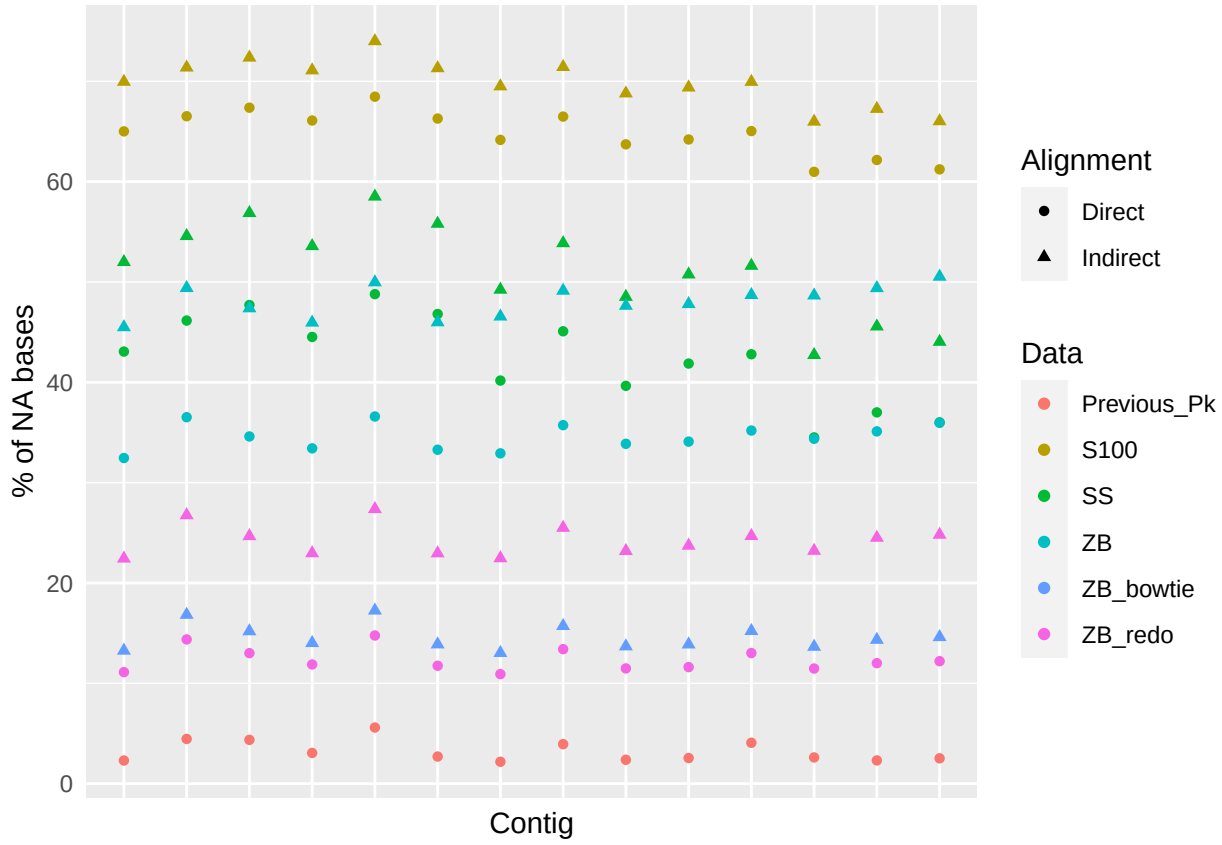


Figure 2: Comparison of read depth across the genome for ZB data

Coloured by alignment:

- To Pk with bwa via removal of human contamination with bowtie2 = light blue.
- Direct to Pk with BWA = red.
- To Pk with bwa via removal of human contamination with bwa = dark blue.

Plot the percentage of bases WITHOUT coverage

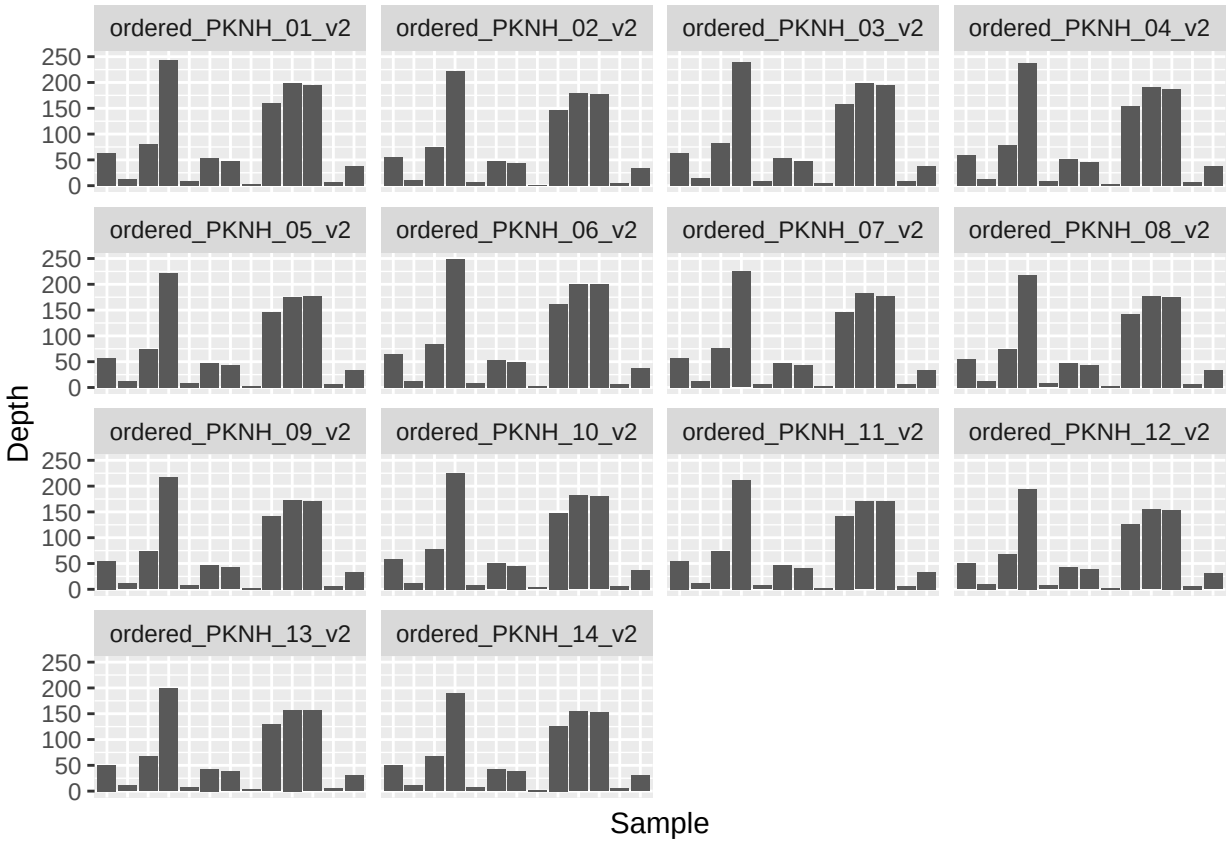


Key points:

- Using ZB's data (ZB_redo) and aligning directly to the Pk genome produces the greatest read depth and is comparable to previous work (Previous_Pk) provided by Ernest.
- Using ZB's data and aligning directly to the Pk genome also produces the lowest percentage of bases without coverage.
- Despite being the best on average, IGV suggests there is inconsistency between samples within the ZB dataset.

Explore the ZB data in more detail

plott the read depth of each sample (direct alignment - ZB) at each contig

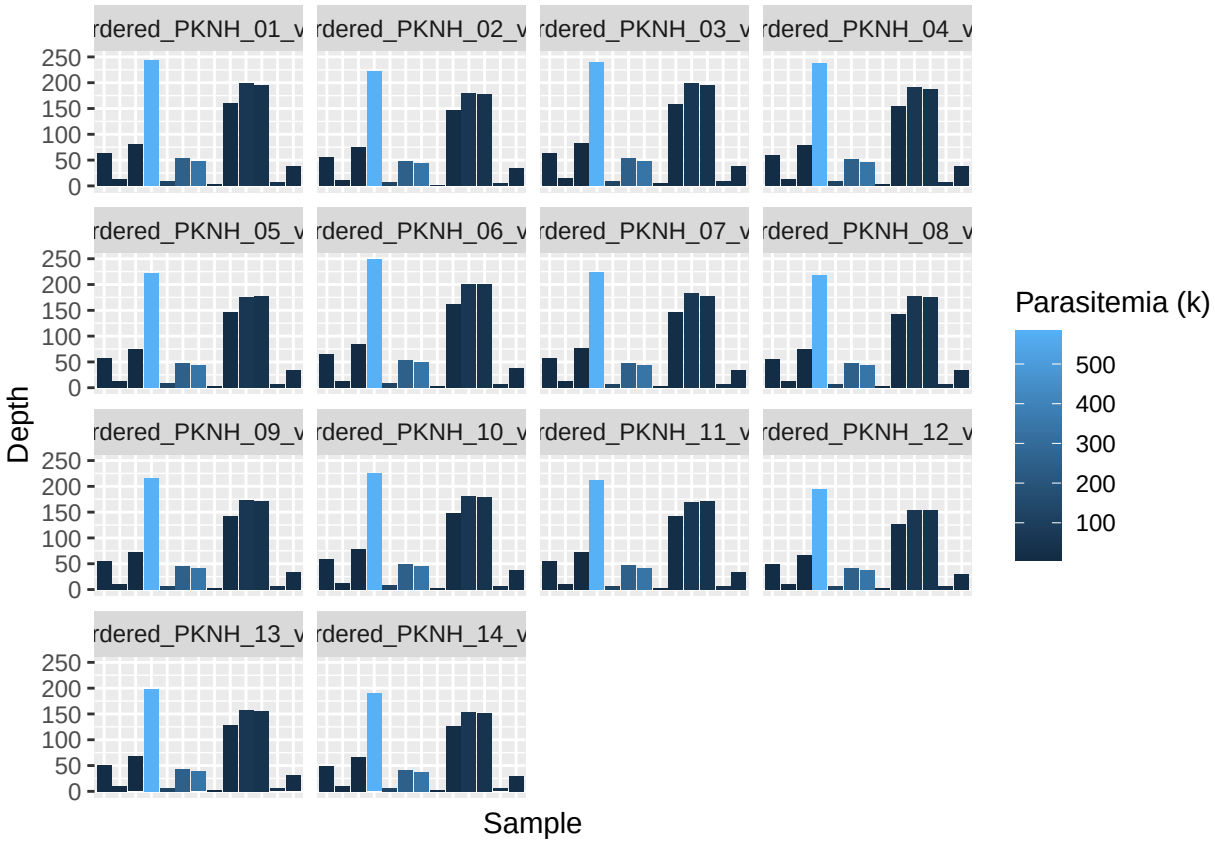


Key points:

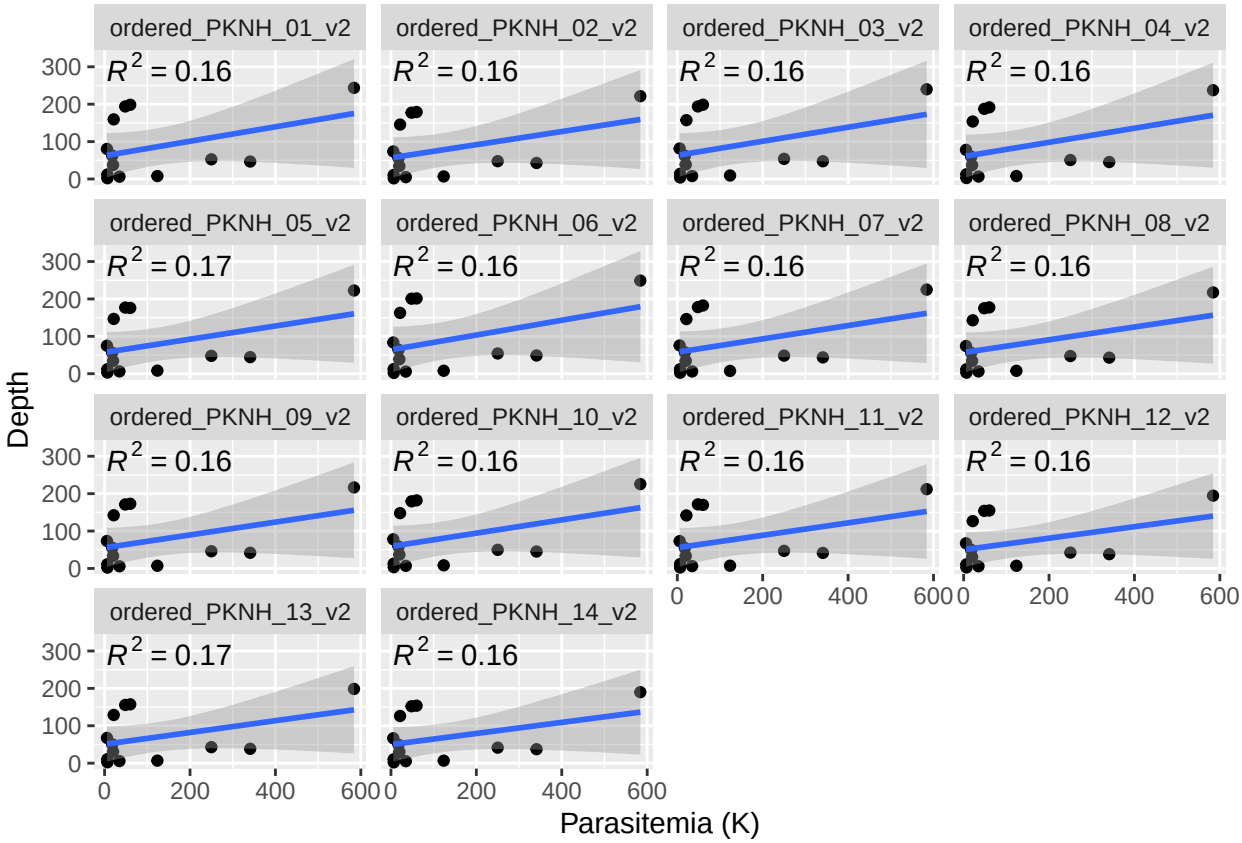
- Despite a great average, the read depth per a sample is inconsistent.
- The read depth across contigs within a sample is consistent.

Read in metadata and combine with summary data to explore the effect of parasitemia alignment

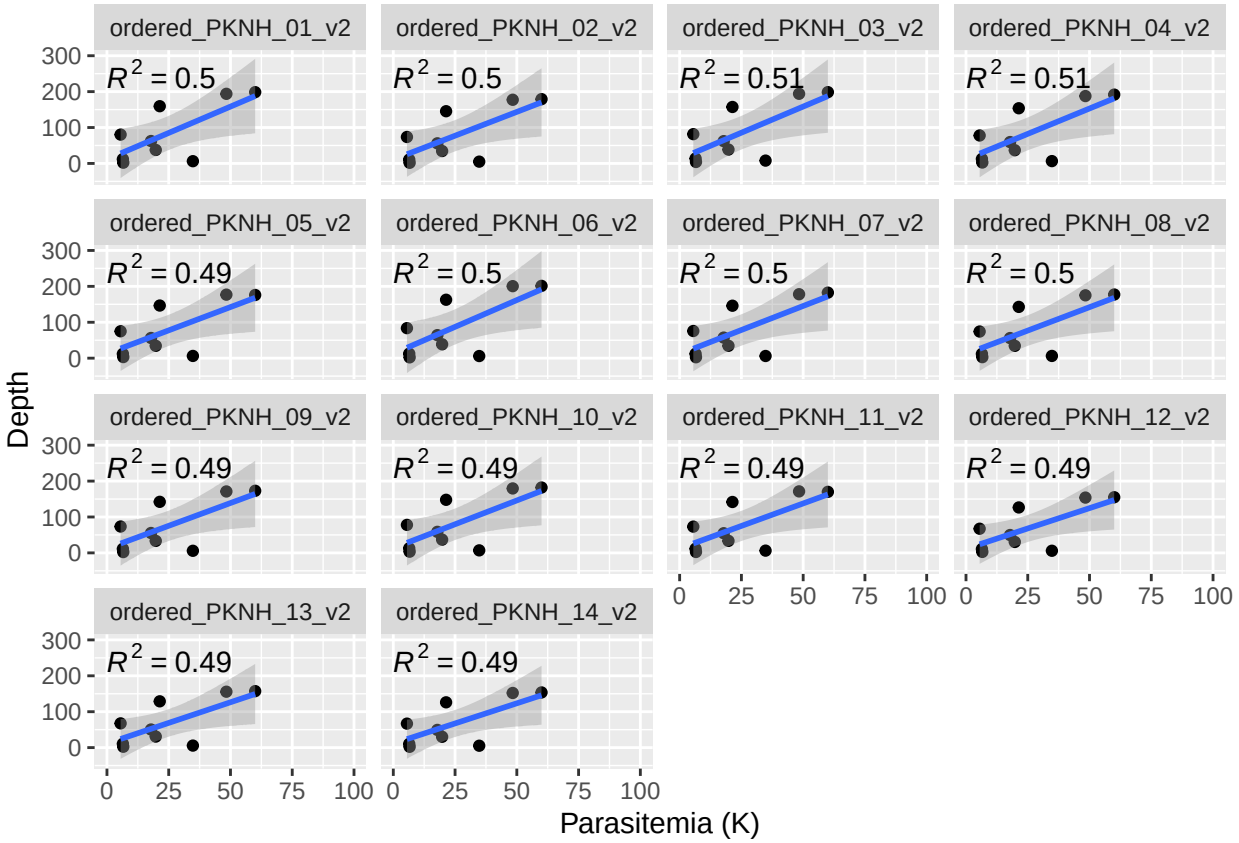
Explore the relationship between parasitemia read depth across samples



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



Plot relationship between depth across samples in relation to parasitemia in the lower parasitemia samples, faceted by contig



Key points:

- There may be a **slight** relationship between paraistemia and read depth, that is stronger at the lower levels of parasitemia.