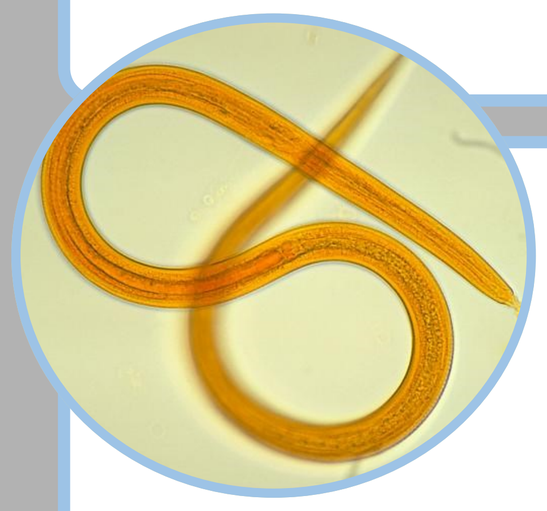
Assessment of long and short-read metagenomics sequencing for the detection of gastrointestinal parasites in faeces

Katie O'Brien[1,2*], Menebere Woubshete[1], Pamela Sachs Nique[1], Jane Younger[1,2], Vicky Hunt[1]

UNIVERSITY OF BATH



Strongyloides stercoralis, an intestinal helminth of humans

Background

- The efficient and reliable detection of parasites is crucial for implementing effective control measures.
- Current diagnostic tools are not fit for purpose, especially where parasitic load is low, indicating the need for alternative approaches.
- Advances in metagenomics sequencing have revealed its potential for non-invasive, untargeted, amplification bias free detection of parasite species in faecal samples.

Aims and Methodology

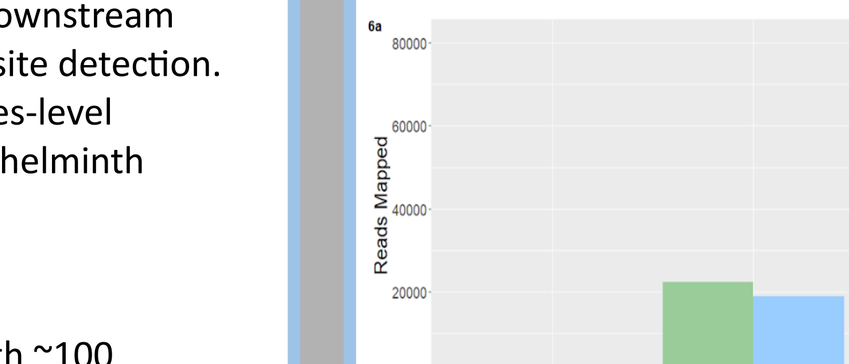
Aim Evaluate long and short-read metagenomics sequencing methods and various downstream bioinformatic analysis tools in parasite detection. In this project, we attempted species-level identification of *Strongyloides ratti* helminth infections from rat faecal samples.

Methods

1) Carry out lab infection of rats with ~100 infective *S. ratti* larvae. Harvest faecal samples from control and infected rats.

2) Extract and sequence DNA using short-read (Illumina) and long-read (Oxford Nanopore) sequencing.

3) Use a range of bioinformatic methods to align DNA sequences to target genome (*S. ratti*) and 13 other nematode species spanning the diversity of nematode clades.



Short-read Alignment Results

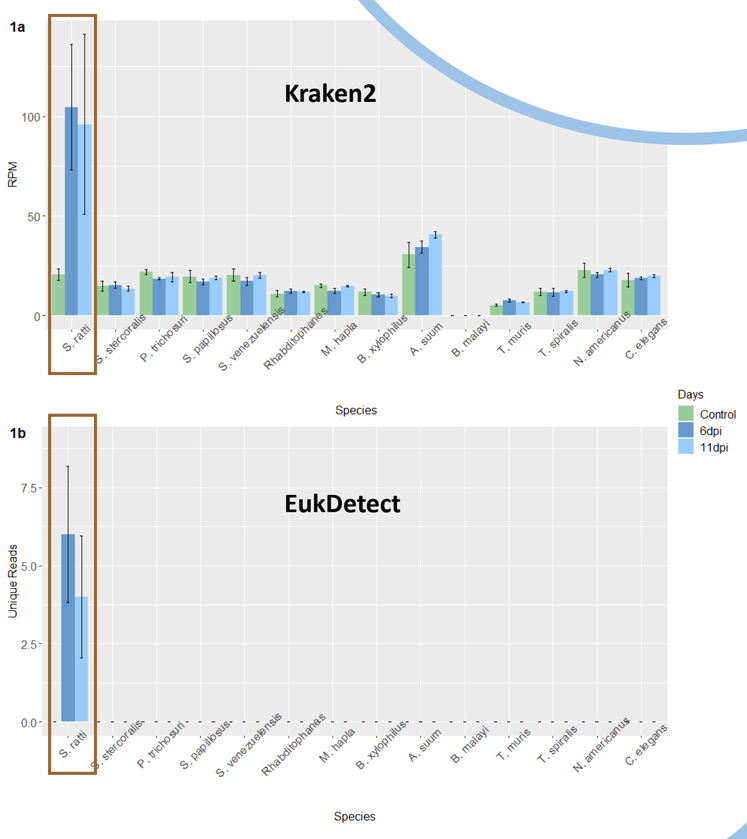


Figure 1: Short-read, shotgun metagenomics DNA reads produced by Illumina, aligned to 14 nematode species using (1a) Kraken2, a broad taxonomic assignment software, and (1b) EukDetect, a species-specific marker gene detection software.

Both Kraken2 and EukDetect showed strong evidence for the presence of *S. ratti.* EukDetect was far more conservative in its assignment but did not produce any false positive alignments to other, non-target genomes or to any reads from the control samples.

RPM = Reads aligned per million reads generated, dpi = days post infection. Three replicates of control, 6dpi and 11dpi were averaged, error bars indicate standard error of average.

Comparative Analysis of Bowtie2 (Short-read) and Minimap2 (Long-read) Alignment Results

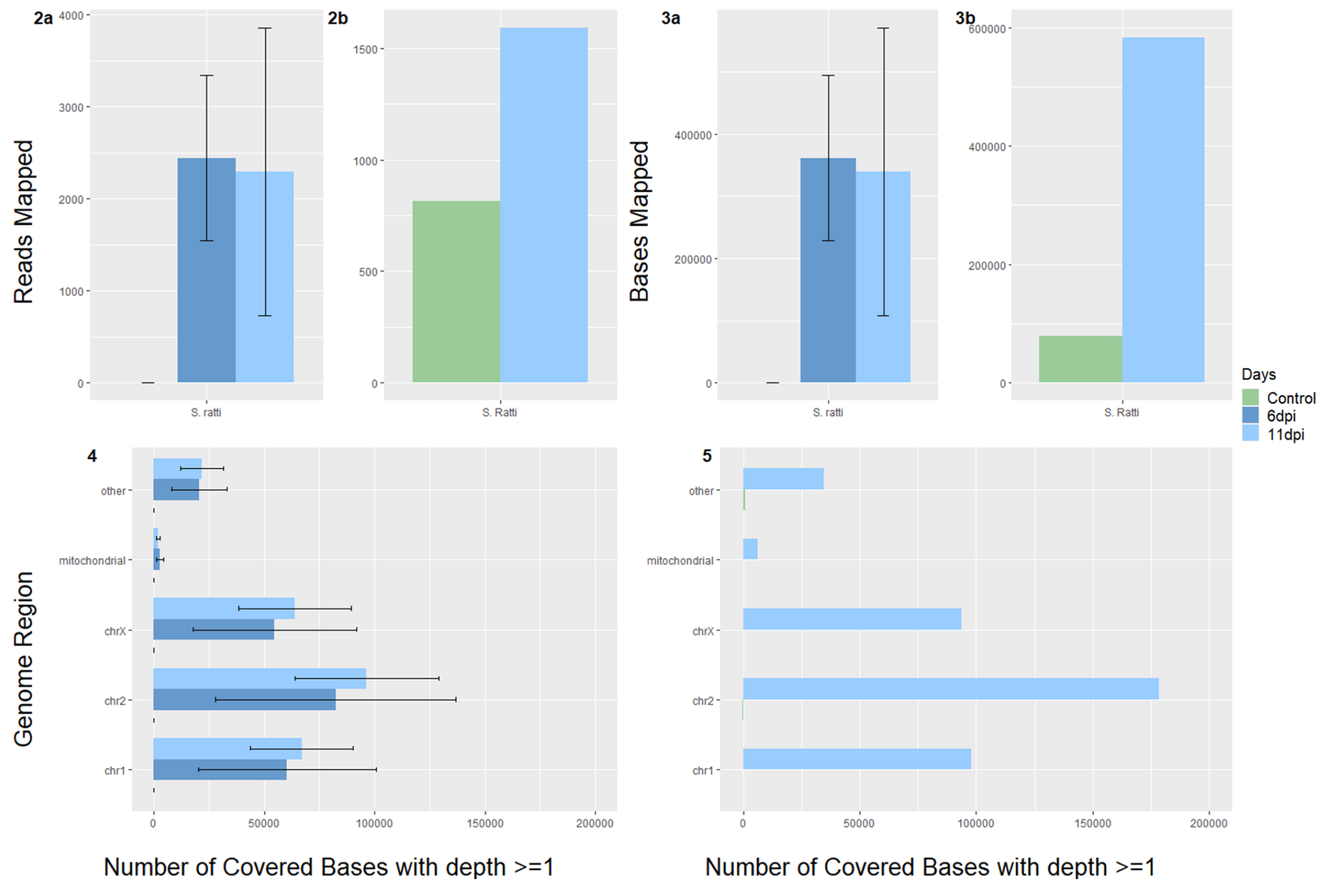


Figure 2: Total number of sequenced reads aligned to the *S. ratti* genome using (a) short-read and (b) long-read sequencing. Error bars indicate standard error. Note that for long-read sequencing, there is no 6dpi sample and no replicates. **Figure 3**: Total number of DNA bases mapped to the *S. ratti* genome using (a) short-read and (b) long-read sequencing. **Figure 4,5**: Total number of DNA bases covered across regions of the *S. ratti* genome using (4) short-read and (5) long-read sequencing. Whilst the number of reads aligning to the *S. ratti* genome is comparable between the two sequencing methods, long-read sequencing generally results in higher overall genome coverage.

Potential Issues in Long-read Sequencing Approach

Figure 6: (a) Long-read, shotgun metagenomics DNA reads produced by Oxford Nanopore, aligned to 3 nematode species using Minimap2. The majority of results produced by Minimap2 were false positive alignments to the *Parastrongyloides trichosuri* and *Brugia malayi* genomes. These alignments were equally present in control and infected samples. In (5b), the root cause of the alignments to the *B. malayi* genome is revealed through analysis of genome coverage. In this case, it is likely that bacterial reads within the sample are aligning to the *Wolbachia* bacterial endosymbiont genome. In (5c), the size of the *P. trichosuri* genome region vs. the number of aligned reads shows that the majority of false positive alignments are to small, poorly constructed regions of the genome. To attempt to solve the issues encountered in long-read sequencing, an attempt was made to use the same marker gene database published by EukDetect on the sequenced reads. Using this method, only one solitary alignment to the *S. ratti* genome was produced, highlighting the lack of appropriate methods for handling long-read data.

Conclusions and Next Steps

This study shows that using broad and species-specific taxonomic assignment software in the analysis of short-read sequencing outputs enables parasitic detection to the species-level. It is shown that while short-read and long-read sequencing produce a similar number of total aligned reads to the *S. ratti* target genome, long-reads generally result in deeper genome coverage, indicating a potential use for long-reads in parasitic helminth genome assembly. Long-read sequencing however, suffers from more error-prone alignments, possibly due to the lack of dedicated long-read tools that act in a similar fashion to programs like Kraken2, or EukDetect. To expand this study further, we are developing new methods of characterising parasitic infection through *de novo* assembly of eukaryotic genomes from metagenomic shotgun sequencing reads. In addition, we are also aiming to create a species-specific marker gene database capable of detecting eukaryotic DNA in error prone, long-read sequencing outputs.

References & Acknowledgements

[1] Dept of Biology and Biochemistry, University of Bath, UK

[2] Milner Centre for Evolution, University of Bath, UK

* kob33@bath.ac.uk

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257 (2019). https://doi.org/10.1186/s13059-019-1891-0

Lind, A.L., Pollard, K.S. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* **9,** 58 (2021). https://doi.org/10.1186/s40168-021-01015-y

NERC GW4+ Doctoral Training Partnership

Milner Centre for EVOLUTION

British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL