Developing Population Genetics Microsatellite Markers from Metagenomic Shotgun Next Generation Sequencing Data

Xavier Barton

Murdoch University: Bachelor of Science – Biomedical Science & Genetics and Molecular Biology

This thesis is presented for the degree of Bachelor of Science Honours, Murdoch University, 2021

Supervisors:

Dr Charlotte Oskam – Senior Lecturer, Murdoch University

Dr Shane Tobe – Senior Lecturer, Murdoch University

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

Xavier Barton

Abstract

Population genetics allows the measure of genetic variation between individuals within a population. Population genetics relies upon the use of genetic markers, which include Amplified fragment length polymorphism, Random amplified polymorphic DNA, Single nucleotide polymorphisms, and Short tandem repeats (microsatellites). Microsatellite markers are an excellent tool for measuring genetic variation due to their high mutation rate and ease of experimentation using polymerase chain reactions. However, the generation of microsatellite markers is costly and complicated. The workflow presented in this thesis aimed at creating an effective microsatellite marker design process. A mixed-species shotgun sequencing sample of an *Ixodes holocyclus* tick was used to create a workflow, with the objective of creating microsatellite population genetic markers. However, as ticks are hematophagous arthropods, the generated shotgun data would also include vertebrate host, microbiome and tick DNA. Therefore, a pipeline was developed, which included read mapping, read classification and metagenomic assembly, to isolate tick genomic content for downstream microsatellite retrieval. Four additional next generation sequencing datasets were obtained from NCBI in order to validate this workflow. In total, read mapping identified 441 microsatellite primer sets, metagenomic assembly identified 2,792, and read classification identified 617 with the un-isolated reads recovering 578 microsatellite primer sets. The results obtained from this study show that performing a Kraken2 custom database classification, read mapping to a close target reference genome (using Bowtie2) and metagenomic assembly (using MEGAHIT) all aid in isolating DNA of a target organism. Overall, the pipeline developed in this study was able to isolate target organism sequences, from which microsatellites were discovered and primer sets generated. The pipeline outlined in this study will provide future researchers a more streamlined approach to creating microsatellite markers for their target organism using standard shotgun next generation sequencing data.

Acknowledgements

Thank you to my friends and family, especially Eliza and Donk, for not letting me worry more than I should have been.

Thank you to Dr Charlotte Oskam and Dr Shane Tobe for being the vital touchstone in guiding me through this project, especially when I had no idea how to articulate my thoughts.

Thank you to everybody at Cryptick Lab for welcoming me to your lab group in the past year and a bit.

Thank you to Siobhon Egan for producing the next generation sequencing data and helping me with some UNIX bits and bobs.

Thank you to Dr David Berryman and Frances Brigg for helping decide on my bioinformatics approaches.

Thank you to Sam for being a wonderful Honours companion and having the perfect Simpsons quote for every occasion.

I also have significant gratitude to the baristas who have made me coffee in the past year.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AFLP – Amplified Fragment Length Polymorphism

AGRF – Australian Genome Research Facility

BAM – Binary Alignment Map

BASH – Bourne Again SHell

bp – Base pairs

BUSCO – Benchmarking Universal Single-Copy Orthologs

BWA (MEM) – Burrow Wheelers Alignment (Maximal Exact Match)

CLI – Command Line Interface

FM-index – Full Text Index in Minute Space

GAGE – Genome Assembly Gold-standard Evaluations

gDNA – Genomic DNA

GUI – Graphical User Interface

HTML – HyperText Markup Language

LCA – Lowest Common Ancestor

NCBI – National Centre for Biotechnology Information

NGS – Next Generation Sequencing

PCR – Polymerase Chain Reaction

QC – Quality Control

QTT – Queensland Tick Typhus

RAM – Random Access Memory

RAPD – Random Amplified Polymorphic DNA

SAM – Sequence Alignment Map

SNP – Single Nucleotide Polymorphism

SRA – Short Read Archive

SSR – Simple Sequence Repeat

STR – Short Tandem Repeat

TBD – Tick-borne disease

Tm – Melt Temperature

WGS – Whole Genome Sequencing

1 - Introduction

The purpose of population genetics studies is to measure and explain the level of genetic variation within and between individuals within populations (Casillas & Barbadilla, 2017). Population genetics can utilise genetic markers that are able to track genetically distinct populations geographically and temporally. In order to use genetic markers as a tool, suitable markers must be discovered, assessed and adapted to use in polymerase chain reaction (PCR) assays. At this time, next generation sequencing (NGS) technologies allow the high throughput of DNA sequence data to be generated in a relatively low amount of time and cost, because of this, being able to develop population genetic markers from this data would be greatly beneficial. As a case study, this project will use Illumina NovaSeq 6000 shotgun sequencing data generated from genomic DNA from a whole body *Ixodes holocyclus* tick, to create a bioinformatics workflow to organise reads from a mixed species NGS sample into their appropriate taxon and will use in-silico methods to identify potential microsatellite markers.

Ixodida (ticks) is an order of hematophagous arthropods that parasitise a wide range of hosts. Ticks threaten the health and wellbeing of animals and humans by their ability to spread infectious diseases. Due to their ability to invoke debilitating allergies and illnesses, several measures have been introduced to reduce the spread of ticks and their pathogens. Ticks are separated into three families, Ixodidae (hard ticks), Argasidae (soft ticks) and Nuttalliellidae (a monotypic family). Of these families, Ixodidae is considered the most medically and veterinary significant with a combined total of 742 Ixodidae species globally (Guglielmone et al., 2020). Australia currently has 74 described tick species and of these the paralysis tick, *Ixodes holocyclus*, is one of great concern due to its ability to induce paralysis in some vertebrate hosts, induce mammalian meat allergy in humans, and for its capacity harbour and transmit tick-borne pathogens (Commins & Platts-Mills, 2013; Graves & Stenos, 2017; Dehhaghi et al., 2019). Understanding the dispersal patterns of ticks, such as *I.*

holocyclus, will allow an understanding of the potential spread of disease and how tick populations may be managed.

1.1 - Next-Generation Sequencing (NGS)

Since its emergence during the mid-2000s, NGS technology also known as high throughput sequencing, has greatly increased in use due to its ability to output large volumes of sequencing data in a relatively short period of time. This itself has allowed whole genomes across many taxa to be fully sequenced, creating applications for medicine (Chen & Zhao, 2019), ecology (Wu et al., 2017), infectious disease (Motro & Moran-Gilad, 2017) to name a few. However, the high data output of NGS technology has unexpected drawbacks as the large quantities of data produced requires significant time for analysis, which is expensive, time-consuming, and computationally taxing (Muir et al., 2016). At this time, bioinformatic analysis rather than the sequencing itself is the bottleneck for many NGS projects. Nevertheless, as more scientists are becoming trained in bioinformatics methods and the cost of computing power is reduced, it is becoming possible for the technology to be greater utilised.

lllumina® (Illumina Inc) is one of the most recognised sequencing platforms in the industry. It provides platforms such as iSeq, MinSeq, MiSeq, NextSeq, HiSeq and NovaSeq, which are short read sequencing platforms, producing reads the are up to 600 bp in length with a relatively high degree of accuracy (Hu et al., 2021). This contrasts to other long read sequencing technologies like PacBio® (Pacific Biosciences of California, Inc), which can generate reads greater than 10,000bp in length but traditionally has had lower accuracy. However, with emerging technologies such as PacBio HiFi sequencing, long read sequencing is become significantly more accurate (Hu et al., 2021). Overall, the NovaSeq 6000 is the most well-rounded sequencing platform that Illumina produces and is able to output 4.8 – 6 Tb of data with 32-40B reads with a read length of 250bp in a single run.

Whole genome shotgun sequencing (WGS) is a method where the entire genome of an organism (or organisms) is sequenced randomly, using short reads of around 100 to 300 bp. Shotgun sequencing

is often used for metagenomic applications where entire microbial communities can be sequenced without having to culture or purify a sample. Using WGS allows relatively fine resolution of taxa down to the species level, which is not possible with amplicon based methods (see below), as well as the data being used in further downstream analysis, with applications such as evaluating phylogenetic diversity, intraspecies polymorphism and gene discovery (Chen & Pachter, 2005; Ranjan et al., 2016). Despite these advantages, WGS sequencing is generally more costly than amplicon sequencing because it requires significantly more data (deeper sequence reads) in order to obtain the required level of coverage for meaningful analysis, however it is still unclear whether WGS sequencing or amplicon sequencing reveals a more accurate biodiversity with some studies revealing less diverse microbe diversity with WGS, compared to amplicon sequencing (Tessler et al., 2017).

Whole genome long read sequencing is emerging as a new standard for sequencing technologies. Long read sequencing directly sequences the DNA molecule, which reduces error brought on by the amplification process of WGS sequencing and produces a longer read length. This aids in repetitive sections of a genome that can be complicated with short read data, as well as the ability to detect structural variations (Huddleston et al., 2014; Pollard et al., 2018).

Another widely used NGS method is targeted amplicon sequencing, which works by using generic PCR primers to amplify variable genes. For example, the ubiquitous bacterial 16S rRNA gene is used for determining bacterial diversity and the internal transcribed spacer (ITS) gene is used for fungi (Buehler et al., 2017; Sperling et al., 2017). The variations in these regions can be analysed and compared to databases, to discover the different genera of bacteria within a sample (Cao et al., 2017). Some advantages of targeted amplicon sequencing is the surplus of bioinformatic tools available for data analysis, the relative low cost compared to WGS sequencing, and the targeted approach results in a much higher coverage for better accuracy (Greay et al., 2018). However, because only one or a few genomic regions are being analysed for genetics variances, a lower level

of taxonomic resolution is provided. This can be amended using WGS where many areas of a genome can be sequenced without bias (Ranjan et al., 2016).

1.2 - Important Steps in Bioinformatics Workflow

To produce meaningful results, an effective bioinformatics workflow must be developed, in order to analyse the substantial amounts of data that are contained in modern NGS datasets. The section below will outline commonly used tools in many bioinformatics workflows. All of the programs described below (apart from named exceptions) are open source, allowing researchers to modify and use them free of charge.

Quality control (QC) is an important first step in any research involving NGS data because the quality of the data that is put into a system will affect the downstream analysis result (Foulkes et al., 2017). Using poor quality data will generate poor final results. Variances in sequence quality can arise in numerous ways. For example, adapter contamination, overrepresented sequences, low quality base calling and read length (Chen et al., 2018). A program that is used extensively for NGS QC is FastQC (Andrews, 2010). FastQC evaluates the intrinsic quality of the FASTQ sequence data in a reliable and efficient way with customisable parameters, outputting a human readable, easily comprehensible HyperText Markup Language (HTML) file with figures demonstrating the quality of the data (Andrews, 2010). A useful extension of FastQC is MultiQC, which utilises FastQC by aggregating data from multiple samples into a single report that is beneficial for visualising results of many samples in an efficient way (Ewels et al., 2016). After the quality of reads has been assessed, poor reads and other contaminants must be removed from the data. A highly cited program for this is Trimmomatic (24,223 citations), which filters and trims poor quality sequence reads and contaminants from the data (Bolger et al., 2014). Trimmomatic is simple to execute, computationally light and reliable. Another effective program for quality control is fastp, which performs all QC steps in one program (Chen et al., 2018). These steps include quality profiling, adapter trimming, read filtering and base correction for short and long read data. Lastly, USEARCH is another commonly used bioinformatics

tool used to explore NGS data at high speeds looking for local and global read alignments from a database (Edgar, 2010). This tool can be used for identifying sequencing controls (such as PhiX control), which may be remaining from the NGS process.

Read mapping is a commonly used technique in bioinformatics that works by aligning reads from an NGS run onto a previously established reference genome, from a database, such as NCBI RefSeq (Pruitt et al., 2007) or a genome assembled in-house. Read mapping is often used for assembling the genome of an organism without having to perform a full de novo assembly. Alternatively, it is used for transcriptomic purposes, for example, differential gene expression (Barrero et al., 2017; Bendele et al., 2019). Two highly used read mapping programs are Bowtie2 and Burrows-Wheeler Aligner (BWA) using the Maximal Exact Match (MEM) algorithm (Langmead & Salzberg, 2012; Li, 2013). Both of these programs are FM-index based aligners (Full-text index in a Minute space) using a Burrow-Wheeler indexing algorithm. They are used to align short read data against relatively large reference genomes because of their low computational load and speed and high sensitivity relative to other mapping programs (Thankaswamy-Kosalai et al., 2017). While Bowtie2 and BWA-MEM are primarily used for aligning DNA sequences, there are also programs, such as Tophat (Trapnell et al., 2009), STAR (Dobin et al., 2013) and HISAT2 (D. Kim et al., 2019), that are primarily used for aligning transcriptomic sequences to a reference genome.

Read classification is a useful tool for assigning reads into categories based on what organism (or more generally taxon) they likely have originated from; this is useful for a mixed species metagenomic samples. Kraken and Kraken2 are popular read binning programs used because of their fast, k-mer based algorithm that use exact matches of a k-mer to a lowest common ancestor (LCA) in the program's database, that results in a high degree of accuracy (Wood & Salzberg, 2014; Ye et al., 2019). Kraken2 works by building an index of all k-mers found in the reference genomes added to the database and then assigns each k-mer to the least common ancestor of all the species that contain the specific k-mer. Then, during classification, Kraken2 matches the k-mers found in the

query sequence reads to this indexed database and eventually assigns the read to the taxon with most matching k-mers by following a path from the root of the taxonomic tree (Wood et al., 2019).

Custom Kraken databases can be built making it useful for binning unique or target species. The databases must be constructed from available genomes, though the genomes do not need to be fully assembled and can be readily found on NCBI. Kraken2 is an updated version of Kraken that used a significantly smaller memory footprint (Wood et al., 2019). The output from Kraken can be visualized using the program Krona, which produces a HTML file with an interactive radial figure allowing for comprehensible data exploration (Ondov et al., 2011). One limitation of Kraken is ambiguous reads will be categorised into increasingly higher taxonomic levels. This can be problematic when there is low genome diversity among sequence reads, meaning that reads will not be classified to their exact species, creating difficulty in estimating species abundance in the sequencing data, as reads from higher levels of taxa are not counted towards individual species. These errors can be corrected by the program Bracken (Bayesian Re-estimation of Abundance with KrakEN) (Lu et al., 2017). Bracken functions by probabilistically re-assigning reads, from the Kraken2 results, into different levels of the taxonomic tree. Reads are distributed from levels higher than the designated taxa level parameter (e.g. the user will set species or genus level abundance estimation) to lower down on the tree to the level that was designated (Lu et al., 2017).

De novo assembly aims to congregate short read sequence data into larger contiguous sequences (contigs) and ideally a full genome in chromosome level structure. A complete de novo assembly is not a time efficient method for many applications, particularly if read mapping or read classification is available and effective, as it is computationally expensive and laborious. However, an initial step of de novo assembly may be effective in order to group some sequence reads together into larger contigs that allow for easier downstream analysis. For example, some studies have shown how a reference guided de novo assembly may be a more practical way of assembling a genome for analysis, even if the genome is not within the same species, rather than starting from scratch

(Lischer & Shimizu, 2017). A popular de novo assembly program is SOAPdenovo2, which is intended to assemble larger genomes (i.e., mammalian genomes), designed for short read data from Illumina sequencing systems, with a minimum memory requirement of about 150Gb (Luo et al., 2012). Another well used program is St. Petersburg Genome Assembler (SPAdes), designed for smaller sized genomes (such as bacteria) and works with Illumina and other short read NGS platforms (i.e., Ion Torrent), using significantly less memory at approximately 9Gb at peak usage (Bankevich et al., 2012). MEGAHIT is a NGS de novo assembler that provides an accurate, conservative assembly of metagenomic sequence data. In terms of performance, MEGAHIT is 3.5x faster than IDBA-UD, another metagenomic assembler and 10x faster than SPAdes, in a test metagenomic dataset (Forouzan et al., 2018; Li et al., 2015). For genomes larger than bacteria, de novo assemblers require significant amounts of memory and storage that many researchers often do not have access to, making de novo assembly a less optimal choice in a pipeline for many projects.

At present, there is no single measurement for assessing the full quality of a genome assembly with researchers constantly debating the best methods, as well as developing new methods to try and produce the most robust genome quality assessment metrics and tools (Thrash et al., 2020).

The N50 score of genome assembly has long been used to determine the quality of a genome assembly, with prior studies noting that it can be somewhat of an indicator of assembly quality, with regard to the size and number of contigs generated. Although commonly used, the N50 score of a genome assembly is now considered an unsatisfactory method (by itself) for assessing the quality of a genome assembly (Thrash et al., 2020). Comparing the N50 scores of datasets of different sizes will not be informative as the datasets (when undergoing assembly) will have a different number and size of contigs generated. As well as this, an N50 score can be inflated due to errors in an assembly, with the score being easily skewed if an assembly has outliers in contig size (Bradnam et al., 2013). It is therefore highly important that when an assembler is being tested for a project the same dataset must be used when determining the best assembler according to the N50 score. This is why,

in conjunction with the N50 score, it is important to use other tools to assess quality and to use tools that measure the particular aspect of a genome assembly that is important to the specific project.

Other genome assembly assessment tools have been developed that aid researchers in determining if a genome assembly would be useful for further bioinformatic analysis and if so, which genomic assemblers to use. Genome Assembly Gold-standard Evaluations (GAGE) is a study of different genome assemblers frequently used. This assessment provides statistics and parameters on which assemblers are best for different projects, along with their advantages and disadvantages (Salzberg et al., 2012). Benchmarking Universal Single-Copy Orthologs (BUSCO) is a tool that allows the assessment of genome assembly completeness using single copy orthologs, to provide a better measure of contiguity than a basic N50 score using evolutionary based gene predictability (Simão et al., 2015). Another commonly used tool is Referee (Thomas & Hahn, 2019), which assesses the quality of a genome assembly based on closely related reference genomes. Referee allows the assessment of both assembly contiguity as well as genome quality, however this method requires the use of reference genomes, which are often not available for many non-model organisms. Using combinations of these assessment tools allows insight into the quality and contiguity of an assembled genome, that a single metric such as N50 cannot provide.

Microsatellite retrieval programs are essential for discovering microsatellites in NGS data. These programs parse the sequence data and display microsatellites sequences along with their flanking regions, length and repeat motif. Many of the once offered programs for retrieving microsatellites are no longer available for use due to lack of documentation and upkeep of download servers. Currently, the most common program that is used is msatcommander (Faircloth, 2008). Msatcommander accepts FASTA formatted sequence files, parses for microsatellite regions, and outputs discovered microsatellites to a CSV file. Msatcommander also has an integrated primer design module to easily create primers from desired microsatellite regions. Another tool is STR detection which is available on Galaxy (www.usegalaxy.org) and scans sequence data for

microsatellites and outputs a tab delimited file with identified microsatellites (Fungtammasan et al., 2015). This program is particularly useful because it is based of Galaxy servers, which allows all the processing to occur on a remote server, making it accessible for any researcher (Afgan et al., 2018). Lastly, Krait is a graphical user interface (GUI) microsatellite retrieval program that is actively maintained and is available on Windows, MacOS and Linux (Du et al., 2018). Krait can parse FASTA files for STRs, compound simple sequence repeats (cSSRs), imperfect simple sequence repeats (iSSRs), minisatellites and macrosatellites, and like other programs has integrated primer design functions. The fact that Krait uses a GUI allows users to identify short tandem repeats (STRs) in datasets with no required knowledge of command line interfaces (CLI). Despite its user-friendly approach, as most other programs are only available on CLI systems like UNIX, parsing data in and out of the program is inefficient.

Primer design programs analyse the flanking regions of a sequence of DNA that needs to be amplified to create a suitable primer set made for PCR analysis. Primer design programs allow researchers to predetermine parameters like target region, primer size, primer melt temperature (*Tm*), as well as many other detailed parameters. Primer3 is one of the most popular primer designing programs currently in use (Untergasser et al., 2012). The benefit of Primer3 is that it is an open-source program that can be used without cost, allowing it to be used locally or using a public web-based Primer3 tool. As Primer3 is open-source, other programs like msatcommander and Krait can have a primer design program integrated, allowing a streamlined marker development system. The software Geneious (Kearse et al., 2012) also includes a primer development feature allowing users to create primer sets for specific sequences. The benefit of Geneious is that it is more user friendly than other programs like Primer3 making primer design more straightforward. However, Geneious is a program that requires a paid licence to use (https://www.geneious.com/).

1.3 - Markers used for Population Genetics

Genetic markers are known points in a genome that can have variability between individuals, allowing experiments to be performed to differentiate individuals within or between populations (Sunnucks, 2000). There are various types of experiments that can be performed for population genetics analysis, each with different strengths and weaknesses. Some of the key marker types that are used include: amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD), short tandem repeat polymorphism (STR / microsatellite), and single nucleotide polymorphisms (SNPs). This section will outline some frequently used genetic markers along with an overview of their strength and weaknesses (summarised in Table 1).

AFLP is a technique by which restriction enzymes cleave a specific site in a DNA sequence, allowing adapter sequences to be attached to the sticky ends. The fragments are then amplified using PCR primers complementary to the adapter sequences. Variation can then be detected by SNPs or insertion-deletion (INDELs) mutations at the restriction sites that prevent cleavage by the restriction enzyme from occurring. The fragments can then be separated by gel electrophoresis with each individual sample producing a different banding pattern depending on their unique mutations (Navajas & Fenton, 2000).

RAPD which uses PCR, but rather than using specific primers to amplify a specific sequence, RAPD uses a multiplex of random PCR primers that are introduced to a large sample of template DNA. The primers then bind to random locations on the DNA and amplify, producing fragments that can be separated using gel electrophoresis. Once the primers are used on different individuals, it will produce a somewhat unique pattern (Butler, 2012; Navajas & Fenton, 2000).

SNPs are a type of point mutation in DNA that replaces one nucleotide in a sequence with another nucleotide, therefore producing biallelic variation. SNP markers are found in the genome by sequencing many individuals in the population and comparing nucleotide changes in these sequences. However, this approach can be a labour-intensive, particularly if there is not a quality reference genome (Helyar et al., 2011; Morin et al., 2004).

STRs, also known as microsatellites, are short DNA motifs (usually less than 8 bp) that repeat many times within the DNA. Microsatellite regions can have a high degree of variability by having a different number of repeats between individuals within a population; this function can be harnessed for population genetics studies. During DNA replication, repeats can be added or lost from the region, and as they do not code for genes there is no adverse effects (Selkoe & Toonen, 2006). The mutation rate of a microsatellite region is relatively higher than coding regions of DNA and is in the scale of around between 10^{-2} and 10^{-6} mutations per locus per generation (Schlötterer, 2000). This means that individuals within a population generally have a high degree of variability in repeat copy number compared to others in the population, making ideal for creating a unique barcode. Though developing microsatellite markers can be somewhat difficult, once markers are available, performing an experiment is simple as it only requires the use of PCR, gel electrophoresis and fragment analysis. The regions surrounding the microsatellite, termed flanking regions, are able to be used for primer development, allowing the microsatellite to be amplified. Though the fast mutation rate provides high resolution and specificity in individuals of populations, the consequence of this is that genetic events that have occurred in the distant past become largely covered up by more recent mutations, unlike markers that operate on a lower mutation rate (Selkoe & Toonen, 2006).

There are a few parameters to consider when choosing a microsatellite for the development of a marker. First is motif length, which is how many base pairs long the unique motif is. For example, the motif length of 'AC' is two and is also known as a di-nucleotide motif, while the motif length of 'AGCCA' is five and is also known as a penta-nucleotide motif. In general, microsatellite motifs with a length of four or greater is recommended, this is because motifs of a shorter length are more unstable and have a greater chance of increasing or decreasing the motif repeat length during PCR amplification, making interpretation of an experiment difficult (Linacre & Tobe, 2013). Secondly, the flanking regions of the microsatellites must be sufficiently unique to create effective primer binding. As the creation of primers is automatically generated in many STR detection programs, such as Krait, microsatellites with flanking regions that are not suitable for primer design are not used. If a

microsatellite detection program does not have a built-in primer design program for the flanking regions (such as 'STR detection' in Galaxy Bioinformatics), primers must be picked manually. This can be done in programs like Geneious, where a flanking sequence can be selected, and a primer pair will be generated. A primer development program implemented into the microsatellite retrieval program is optimal as it reduces the time required to import and export data in and out of several programs. Finally, microsatellite length will determine how many repeats of a motif there will be in the microsatellite. There are no specific requirements for choosing microsatellite length as this will be varied depending on the population that the marker is being used in. However, microsatellite length that is too short may be challenging to interpret when performing the genotyping experiment.

Table 1. Overview of Genetic Markers Frequently Utilised for Population Genetics Studies and their Advantages and Disadvantages (Butler, 2012; Helyar et al., 2011; Morin et al., 2004; Narayanan, 1991; Navajas & Fenton, 2000; Selkoe & Toonen, 2006).

Marker Type	Description	Advantages	Disadvantages
Amplified Fragment Length Polymorphism (AFLP)	Restriction enzymes are used on a sample, variation in SNPs and INDELs may prevent restriction enzymes in bind creating variation in fragment number and length.	- Relatively reproducible - No marker development - Inexpensive - Prior knowledge on sequence is not needed - Distinguishes closely related populations	- High quality DNA required for enzymes to bind correctly - SNPs and INDELs are generally only biallelic - Can be difficult to interpret - Difficult to distinguish heterozygotes from homozygotes
Random Amplified Polymorphic DNA (RAPD)	Selected PCR primers amplify random sequences of the template DNA. Mutations between individuals will affect how primers bind, resulting in different amplification products being produced.	- No knowledge of target DNA is required - Results are generated quickly - Able to design primers for specific regions from the results - Inexpensive	- Works poorly on degraded DNA - Lower resolution than other available methods - May have poor reproducibility - Only displays dominance - Can be difficult to interpret
Short Tandem Repeat Polymorphism (STR / Microsatellite)	A segment of repetitive DNA which has a high mutation rate. The differences in size can be analysed between individuals.	- Highly polymorphic - Once markers are developed the experiment is easy - Highly reproducible - Widely used	- Marker development is labour intensive - Can be very species specific - Poor for looking at events in distant past
Single Nucleotide Polymorphisms (SNPs)	A point mutation that changes a single nucleotide for another nucleotide at a specific point.	- Very abundant in genome - Can provide a more representative of variance in the genome - Easily applied to high throughput sequencing - Good for poor quality samples - Simple mutation model	- Only biallelic so there is less information than STRs - Many SNPs are required for a statistically significant result - Susceptible to ascertainment bias

1.4 - Ticks

Ixodida (ticks) are obligate hematophagous arthropod ectoparasites categorised within the class Arachnida, a group of joint legged arthropods which include Araneae (spiders), Sarcoptiformes (mites) and Scorpiones (scorpions). Most ticks (excluding the single species within the Nuttalliellidae family in Africa) belong to either the Ixodidae (hard tick) family or Argasidae (soft tick) family. Both hard and soft ticks are known to parasitise a vast number of vertebrates, including reptiles,

amphibians, mammals, and birds (Parola & Raoult, 2001; Kwak & Madden, 2017; Mendoza-Roldan et al., 2020). Ticks are located on every continent (Barbosa et al., 2011) as they can live in a diverse range of habitats, making them an important organism in many regions from an ecological perspective (Jongejan & Uilenberg, 2004). Most adult tick species are around 3-5mm in length (unfed), of brown-red colour, with eight jointed legs connected to a fused abdomen. Their capitulum contains the mouthparts that allow the tick to puncture and hook into the skin and draw blood (Sonenshine & Roe, 2013).

Ticks require a blood meal at each of the three-life stages (larva, nymph and adult). However, differences are observed between tick species where either one, two or three vertebrate host species are required (Labruna et al., 1997; Sonenshine & Roe, 2013). In a three-host life cycle, larva feed for the first time on a host, then detach into the environment to develop for several weeks. The larva then moults into a nymph which feeds for a second time on another host. The nymph then metamorphosises into either an adult female or male tick. At the adult stage, the female tick will mate with a male tick, feed for a third time, and will produce between 2000 and 20,000 eggs and then dies. The male ticks will continue to mate with more female ticks, while only briefly feeding (Barker & Walker, 2014; Jongejan & Uilenberg, 2004).

Ixodida are first identified in the fossil record in the Cretaceous period, around 65 – 146 million years ago (de la Fuente, 2003). Hard tick species are often considered the most important group of ticks as they have the widest variety of species (around 683 globally) and spread a broad range of diseases to their vertebrate hosts. Hard ticks are characterised most easily by hard plates, called a conscutum in males or scutum in females on the dorsal surface of their bodies and when viewed dorsally, the mouthparts of hard ticks are anterior projecting (Jongejan & Uilenberg, 2004). Soft ticks are the other main family of ticks containing approximately 183 species (Jongejan & Uilenberg, 2004). These ticks are distinguishable from ixodid ticks by the lack of a hard plate on the dorsal surface as well as mouthparts not being visible when viewed dorsally (Horak et al., 2003). In Australia, there are

currently 74 species of ticks; 14 argasid and 60 ixodid tick species that have been discovered on a wide variety of vertebrate hosts (Barker, 2019; Evans et al., 2019). For the purpose of this study, the remainder of this review will focus on hard ticks (herein referred to simply as ticks).

Ticks cause major disruptions to the health sector, wildlife conservation and the livestock industry making them a key target for research. In the livestock industry, production loss occurs because of tick infestations due to resistance to acaricides and tick-borne diseases (TBD). Surveys conducted by Playford (2005) show that an estimated $170-200m AUD is lost due to tick infestations in the cattle industry (Playford, 2005).To reduce this heavy burden significant biosecurity measures are taken to enforce tick borders around Australia, which in turn help protect wildlife and livestock, and limit the spread of TBD. These practices include the NSW/QLD tick boarder, which provides a tick control point to prevent tick infestations between NSW and Queensland, therefore limiting the number of wildlife and livestock affected by ticks and their diseases. However, this border costs local government an estimate of $3.1 million AUD per year to maintain, with no large scale studies showing how effective these control measures are (Chudleigh & Franco-Dixon, 2010). In addition to the monetary loss associated with defending geographical spread of ticks, non-native introduced ticks can cause detrimental effects to native Australian wildlife, due to their parasitic nature and the ability to cause and spread diseases that the native wildlife may not have evolved resistance to. As well as wildlife, companion animals are at great risk of both native and introduced tick infestations with one survey examining 4,765 ticks removed from 837 companion animals including dogs, cats, and horses. Of these ticks 11 species were identified, including ticks that are known to cause disease (Greay et al., 2016). Finally, in the health sector, a tick's ability to carry pathogens and trigger disease is becoming more widely recognised and is increasingly alarming.

1.5 - *Ixodes holocyclus* and Associated Diseases

First described in 1899 by L.G. Neumann, *Ixodes holocyclus*, also known as the paralysis tick, is one of the most medically and veterinary significant ticks in Australia, due to its disease causing capabilities

(Neumann, 1899). This species has an enzootic range along coastal regions of eastern Australia from Northern Queensland to Bairnsdale, Victoria and is known to inhabit wet and dry forested areas (Barker et al., 2014; Jackson et al., 2007). While adult *I. holocyclus* ticks can be easily identified (due to pigmented 1st and 4th coxae/legs), *I. holocyclus* cannot easily be morphologically distinguished from *Ixodes cornuatus* and *Ixodes myrmecobii* in juvenile ticks or ticks with damaged morphological features (Evans, 2018). Therefore genetic testing is required for an accurate identification, whereby genetic differences can be observed by comparing the, COX1, ITS2 mitochondrial and nuclear genes (Song et al., 2011). *Ixodes holocyclus* is a three host species and has been reported to parasitise around 34 mammalian and avian species; however, the primary host is hypothesised to be bandicoots (*Isodoon macrourus* and *Perameles nasuta*), although it is the most common human biting tick on the east coast of Australia (Barker et al., 2014, p. 70).

Tick paralysis is a disease caused by envenomation of toxins in tick saliva that result in paralysis of the vertebrate host. In 1976, B. J Cooper and I. Spence published a study that described how ticks secret a neurotoxin (holocyclotoxin) that inhibited the release of acetylcholine from neuromuscular junctions, which in turn leads to muscle paralysis (Cooper & Spence, 1976). Later it was observed that a single tick can produce enough toxins in order to induce paralysis in the vertebrate host; clinical signs generally present from three or four days of a tick feeding (Chand et al., 2016). Located in Australia, *I. holocyclus* is one of the most recognised ticks to cause this disease as it appears to have a high level of virulence compared with ticks found outside Australia that are known to cause paralysis, such as *Dermacentor andersoni* and *Dermacentor variabilis* in North America and *Ixodes rubicundus* and *Rhipicephalus evertis* in Africa (Masina & Broady, 1999). *Ixodes holocyclus* has been shown to cause tick paralysis in both humans and wildlife in Australia. In humans, symptoms can range from facial palsy to whole body flaccid muscle paralysis and, though now rare, death from respiratory failure. Children (1-5 years) are particularly susceptible to the onset of tick paralysis (Grattan-Smith et al., 1997; Pek et al., 2016).

Though wildlife is rarely affected by native ticks, due coevolution, some occurrences have been observed. An example from wildlife is the deaths of *Pteropus conspicillatus* (Spectacled flying-foxes) that has resulted from paralysis induced by *I. holocyclus* in Northern Queensland. As *P. conspicillatus* is a threatened species, any negative impact on the populations of these flying foxes needs to be regarded with concern (Buettner et al., 2013). Another study showed the effects of three ticks (*Haemaphysalis humerosa, Ixodes tasmani and Ixodes holocyclus*) on *Isoodon macrourus* (Northern Brown Bandicoot) with a reduced body weight increase and increased leukocyte count (Gemmell et al., 1991). As well as wildlife, companion animals are known to be at risk of tick paralysis, with 3479 canine and feline cases being reported by veterinarians between the 1st of September 2010 and the 31st of January 2012 (Eppleston et al., 2013).

Alpha-gal Allergy, which is also known as the Mammalian Meat Allergy, is the development of an allergic reaction to consuming mammalian meat or sometimes products made from mammalian meat such as gelatine. In Australia, the allergy is caused by the tick *I. holocyclus*, when this tick bites it can transfer a carbohydrate allergen (galactose-alpha-1,3-galactose) to the host, generating sensitivity to the carbohydrate (van Nunen, 2015). Galactose-alpha-1,3-galactose (alpha-gal) is not present in primates but is commonly found in relatively large quantities in the blood of other mammals, so the carbohydrate is taken up from the blood of a previous non-primate host and is later delivered into the human host via a second blood meal (Commins & Platts-Mills, 2013). This, in turn will sensitise the human host to alpha-gal and the host's immune system will generate large levels of IgE to combat the allergen, generating an allergic reaction. Symptoms of the allergic reaction occur around three to six hours after consuming mammalian meat and include nausea, diarrhea, indigestion, hives, angioedema and anaphylaxis (Wong & Sebaratnam, 2018).

In recent years, the paralysis tick has been the centre of debate about the presence/absence of a northern hemisphere TBD (*Borrelia burgdorferi*) (Irwin et al., 2017; Piesman & Stone, 1991). However, *I. holocyclus* is not a competent vector but given its proximity to many people complaining

of TBD, further research is currently underway to determine which, if any, microbes are contributing to this unknown illness in people.

Queensland Tick Typhus (QTT) also known as Australian Tick Typhus is one of two recognised rickettsial type diseases that occur in Australia; the second is Flinders Island spotted fever (*Rickettsia honei*) associated with the reptile tick, *Bothriocroton hydrosauri*. QTT is caused by the bacteria *Rickettsia australis* and is predominantly found on the east coast of Australia. The disease is known to be transmitted mainly by *I. holocyclus* and can be transmitted either vertically between the ticks life stages (transstadially) or horizontally in a cycle between tick and its vertebrate host (Dehhaghi et al., 2019; Unsworth et al., 2007). Symptoms of QTT generally appear after four to ten days after a tick has attached to the host and include eschar around the bite area, fatigue, fever, headache, myalgia, and a maculopapular or vesicular rash that follows within ten days (Parola et al., 2013).

Q Fever is a zoonotic disease caused by the bacteria *Coxiella burnetii*. The bacteria are extremely resistant to many kinds of stress, it is able to tolerate high levels of heat, chemicals, digestive enzymes and antibiotics. The bacteria are able to resist this due to its ability to form spores that can survive for long periods of time (Duron et al., 2015). *Coxiella burnetii* is primarily spread through airborne particles when infected livestock are present, but it is thought that ticks may act as vectors for the bacteria. The consensus on whether ticks (including *I. holocyclus*) transmit Q fever are not yet confirmed, however there is some strong evidence that supports the case that it does. Data from several studies suggest that there can be high levels of *Coxiella* bacteria in ticks collected from various animals (Loftis et al., 2006; Cooper et al., 2012; Pacheco et al., 2013; Graves et al., 2016). In addition to this, several papers describe experimental conditions that have been able to show the ability for a tick to uptake the *C. burnetii* from an infected animal and transmit it to an uninfected animal (Derrick et al., 1942; Široký et al., 2010). In spite of this, transmission of *C. burnetii* from a tick to an uninfected host in a natural environment has not be confirmed and many of these studies took place in the 1940's with little current experimentation.

1.6 - Next-Generation Sequencing of Ticks

Identification of Microbes

The primary application of NGS in ticks is currently used for the analysis of the bacterial communities within ticks (referred to as microbiome), due to its high level of sensitivity compared to traditional methods, such as culture and Sanger sequencing. Analysing the microbiome provides important insights into ticks, including their feeding patterns or host preference, life cycle, environment, general ecology, as well as identification of medically and veterinary significant pathogens (Greay et al., 2018). With an ever-increasing recognition of the role that endosymbionts and pathogens have within the tick microbiome and the potential influences this can have on humans, domestic animals, livestock and wildlife, microbiome NGS analysis is becoming increasingly important. Using this approach Brinkmann et al. (2019) found the presence of *Rickettsia, Coxiella, Francisella, Borrelia, Babesia, Theileria* and *Hemolivia* as well as endosymbionts associated with *Coxiella* and *Francisella* in questing and feeding ticks collected from several locations around Turkey, showing the scope and sensitivity that can be provided when using NGS technology in this area, allowing future tracking of these pathogens (Brinkmann et al., 2019). Using NGS 16S rRNA amplicon sequencing Egan et. al. was able to highlight the high levels of diversity in bacteria contained within the microbiome of native bandicoot ticks, discovering novel species of *Coxiella*, *Ehrlichia*, *Francisella*, *Neoehrlichia*, and *Rickettsiella*. Studies like this, will allow the continued surveillance of wildlife microorganisms, providing advancements in pathogen detection and without the use of NGS technology the required specificity for detecting these novel microbes would not be available (Egan et al., 2020).

Tick Genetics

To the authors' knowledge, at the present time a fully assembled tick reference genome is not available to the public. However, several draft assemblies are available and have been generated in addition to numerous completed mitochondrial genomes of ticks. Currently, the most developed tick genome is of *Rhipicephalus microplus* [NCBI Genome: 2797]. Much of this genome is sorted into

chromosome level structure that makes it reliable and useful for analysis as the genome is more advanced in its assembly. The remaining preliminary tick genome assemblies (acquired from i5k - http://i5k.github.io/) include: *Ixodes ricinus* [NCBI Genome: 16267] (present only in Europe), *Ixodes scapularis* [NCBI Genome: 523] (present only in North America) and *Haemaphysalis longicornis* [NCBI Genome: 69202] (globally distributed). These genomes have revealed the genome of mites are unexpectedly small compared to the exceedingly large genomes of ticks (Gregory & Young, 2020). In addition, the transcriptome of ticks has revealed variance of gene expression of adult feeding ticks. This study found potential differences in acetylcholinesterase-like contigs, peptidase, proteases and protease inhibitors between fed and unfed ticks, which may have impacts of future studies on the tick feeding cycle (Bendele et al., 2019). These studies show that access to genomes constructed from NGS data allows investigation that were not previously possible due to the lack of bioinformatic computing power and the technological limitations of Sanger sequencing.

The widespread and unrestrained use of acaricides has resulted in the development of resistance to these products in tick populations, leading to the greater demand for an effective tick vaccine for cattle (Yessinou et al., 2016). Although there are tick vaccines on the market, many have been available for over 20 years and have varying levels of efficacy. The large growing body of new tick vaccine research is utilising NGS technology to discover effective antigens for vaccine development. One study by Guerrero et al. (2014) found that the aquaporin protein RmAQP1 may work as an effective antigen in a *R. microplus* vaccine, creating a 75% and 68% efficacy in two different trials (Guerrero et al., 2014). Subsequently, a novel study using a vaccine including a recombinant trypsin inhibitor from the Kunitz-BPTI family of *R. microplus*, yielded a 32% efficacy against *R. microplus* (Andreotti et al., 2012). Without the genomes provided by NGS technology the potential vaccine proteins discovered by these studies would not have been found, demonstrating the systematic use of NGS in many studies.

1.7 - Population Genetics Applications for Ticks and the Markers Used

There are three primary reasons researchers have employed population genetics on ticks. The first of which is understanding the geographical dispersal of different tick populations, this allows the analysis of how far different tick populations are spreading in a region with variations such as host mobility and seasons. Secondly, tracking different populations and sub-populations of ticks allows the tracking of TDB, with the potential that different populations may carry a unique range of pathogenic species. Finally, the biosecurity risks posed by different tick populations must be measured to keep wildlife and livestock safe from native and introduced ticks. The section below will outline some of the most frequently used population genetic marker types used with applications in tick research.

AFLP

The method has been used in a paper studying mites by Weeks et al. (2000), the study was able to use AFLP analysis to highlight how genetic variation of mites can be assessed using the AFLP technique. Weeks et al. used AFLP because of its ability to screen DNA regions throughout the genome and the ability to produce distinct genetic patterns without the need for prior genome knowledge. However, it is noted that this technique has difficulty in distinguishing homozygotes from heterozygotes. Another example was the development of 335 AFLP loci from 750 Tsetse flies, whereby two genetic populations of the flies were identified in Zimbabwe. In this study, AFLP was employed for its accessible and reliable procedure to distinguish closely related populations (Lall et al., 2010). Finally, Araya-Anchetta et al. (2013) used AFLP markers to create a greater understanding of natural hybridization events in *Dermacentor andersoni* and *Dermacentor variabilis* tick species in North America (Araya-Anchetta et al., 2013).

RAPD

RAPD markers have been used for their power in distinct profiling patterns, as well as the small number of primers required to produce highly polymorphic markers. This enabled a simple and

reliable way to determine genetic distances between *I. ricinus* populations in Lithuania (Radzijevskaja et al., 2005). Another example in which the RAPD technique was used, was to show how different strains of *R. microplus* had varying band patterns which were related to either acaricide-susceptible or acaricide-resistant ticks (Hernandez et al., 1998).

SNPs

Determining the population structure of organisms is a common application of SNP analysis, for example a paper by Poli et al. (2020) sampled 497 *I. ricinus* ticks from 28 different tick populations across 20 countries in Europe and North Africa and developed 125 SNP markers. With these markers the researchers determined that the population structure of the ticks was very strong and found a population separation between the European and North African ticks. SNP markers where used for this research because it is possible to create a large bank of reliable markers suitable for population genetics in a short amount of time, however it is noted that SNPs have a much lower mutation rate compared to other marker types (Poli et al., 2020). Another study by Nadonly et al. (2015) used SNP markers from the mitochondrial genome of *Ixodes affinis* and *Amblyomma maculatum* (which share many vertebrate hosts), to avoid the process of developing more time consuming species-specific microsatellite markers. This study collected ticks from various locations around North America and unexpectedly found that the population structures of each of the species were unique going against the hypothesis that the populations are spreading north. As well as this their evidence shows discrete geographical populations of *A. maculatum* from *I. affinis* and that *I. affinis* was comprised of two genetically differentiated populations (Nadolny et al., 2015).

STRs

A paper by Talbot et al. (2020) is using microsatellite markers to assess the invasion of *Ixodes scapularis* into Canada. The study consisted of collecting between three and ten *I. scapularis* samples from each site at various locations around Canada and used previously established microsatellite markers from Fagerberg et al. (2001). These types of markers were used in this study because of the

highly polymorphic nature, as well as other studies in the area that utilised microsatellite markers. Talbot et al. suggested that a downside of using microsatellite markers for this study was the limited number of markers across the genome compared to a marker like SNPs. After statistical analysis, the study found results that supported the hypothesis that tick populations were being carried in long-distance movements possibly by birds rather than a slow spread causes by terrestrial hosts (Fagerberg et al., 2001; Talbot et al., 2020). Guzinski et al. (2008) gives an example of the development of microsatellite markers. The study created these markers to analyse the population structure of the Australian reptile tick *Bothriocroton hydrosauri*, allowing these markers to be used for any future population genetic studies of the species (Guzinski et al., 2008). Microsatellite markers may also be used to test paternity, this has been demonstrated in a paper by Cutullé et al. (2010) who used microsatellite markers, because of their highly polymorphic nature, useful for testing paternity, to determine if *R. microplus* ticks where the result of mating with one or more mates. The experiment used 15 fed female tick removed from three different cows in Brisbane, Australia and performed DNA extraction and microsatellite amplification using markers previously developed by Cutullé et al. (2009). The study confirmed that there was multiple paternity in over 66% of the sampled ticks (Cutullé et al., 2009, 2010).

Tick population genetic markers are extremely beneficial for the tracking of populations around areas involved in the livestock industry, due to the detrimental effects ticks can cause. Busch et al. (2014) provides an example of this method, where the researchers aimed to track the movements of *R. microplus* in southern Texas on infested livestock. This was completed by first sampling ticks from cattle and deer in southern Texas, then comparative analysing on SNP and microsatellite makers for different species of tick. Using this approach, Busch et al. observed *R. microplus* remained a persistent infestation for multiple generations at a location with four distinct genetic populations. This information was then used to track the dispersal of these distinct groups due to cattle transportation (Busch et al., 2014).

1.8 - Thesis Aims and Hypothesis

Once microsatellite markers have been created, they are an excellent tool for analysing genetic variation within populations. However, the process of developing population genetic markers is often time consuming and complicated. The creation of a bioinformatics workflow to quickly produce microsatellite population genetic markers would be greatly beneficial. This is especially true with ticks, whereby tracking their population has significant medical, veterinary and biosecurity importance.

The research hypothesis of this study is based on microsatellite markers previously developed for *I. scapularis* and *R. microplus* (Fagerberg et al., 2001; Busch et al., 2014), whereby it is hypothesised that by mining *Ixodes holocyclus* NGS data, STRs will be identified that are suitable for population genetics applications in Australian *Ixodes* ticks.

The research aims of this thesis are three-fold: first, to confirm the quality and usability of the NGS data for potential downstream analysis; second, to organise reads from metagenomics dataset into their appropriate taxonomic classification; and finally, to use an in-silico method to identify potential microsatellite markers from organised Ixodida shotgun NGS reads.

2 - Materials and Methods

This project will utilise NGS data from an Illumina Novaseq 6000 of a mixed tick sample that includes *I. holocyclus* (paralysis tick), bandicoot DNA (the tick host) and DNA from the microbiome of the tick that will include a whole range of bacteria, viruses and protozoan parasites. Due to this, the tick DNA will need to be isolated from other taxa DNA. The general pipeline that this project will undergo will be as follows (Figure 1). Quality Control that will assess the quality of the raw reads that come from the sequencing machine. The next programs will be used to trim and remove poor quality reads that were detected from the quality control step. After this, the data can take three directions, the reads can be binned, meaning placing them in taxa or the reads can be mapped to a reference genome to isolate reads specific to it or alternatively a de novo assembly can be performed to start generating contigs out of the sequence reads. After the sequence reads are assembled in some way the data can be run through a program that will search for microsatellites. The final step in this pipeline is to develop primers for the best microsatellite markers that will work for PCR analysis. Once these steps have been completed wet lab testing of the potential markers must be performed.

2.1 - Sample Collection and DNA Extraction

An adult female *Ixodes holocyclus* tick used for sequencing was collected in Boorie Creek, New South Wales, Australia, on the 6th of September 2014. The tick was parasitising a bandicoot and was removed and stored in 70% ethanol. Genomic DNA (gDNA) was extracted from the entire tick, including the bloodmeal and associated microbes using the QIAGEN DNeasy Blood and Tissue Kit (QIAGEN, Germany) following the manufacturers supplementary protocol “Purification of total DNA from insects using the DNeasy Blood & Tissue Kit”. Extractions were performed by Siobhon Egan. The extracted DNA was stored at -20 °C for future use.

2.2 Sequencing

The extracted gDNA obtained from *I. holocyclus* was subjected to paired-end shotgun sequencing using the Illumina NovaSeq 6000 platform performed by AGRF, using the Illumina Nextera XT DNA library kit. The sequencing run analysed in this project was conducted over a single lane, with the lane (Lane 1) yielding 10,022,726 x 250bp paired-end reads (5.03 Gb) provided in the FASTQ format. Initial image analysis was performed in real-time using the NovaSeq Control Software v1.6.0 and Real Time Analysis v3.4.4. After this, sequence data was generated using the Illumina bcl2fastq 2.20.0.422 pipeline.

2.3 - NCBI Datasets

Alternative datasets were downloaded from the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA). These datasets included *R. linnaei* (SRA: SRR13348027)*, R. appendiculatus* (SRA: SRR11263445)*, D. variabilis* (SRA: SRR5317834) and Tick Virome (SRA: SRR13164901). All of these datasets were Illumina shotgun datasets of various tick species. All bioinformatic experiments performed on the original AGRF dataset (Lane 1) were also performed on the additional NCBI datasets to develop and validate the sequence read isolation workflow.

Rhipicephalus linnaei

The dataset titled 'WGS of *Rhipicephalus linnaei*: whole adult' (formally *Rhipicephalus sanguineus*) (Šlapeta et al., 2021) is an Illumina HiSeq 2500 paired-end whole body shotgun sequencing run on the tropical brown dog tick (*R. linnaei*) submitted to NCBI by the University of Sydney on the 4th of January 2021. The compressed 1.2Gb dataset comprises a single forward and reverse FASTQ file containing 12,619,829 x 150bp long reads totalling 1,892,974,350bp each.

Rhipicephalus appendiculatus

The dataset titled 'WGS of *Rhipicephalus appendiculatus*: semi-engorged adult female' is an Illumina HiSeq 2500 paired-end whole body shotgun sequencing run on a brown ear tick (*R. appendiculatus*)

from Kenya and was submitted to NCBI by the University of Sydney on the 8th of March 2020. The compressed 753.7Mb dataset comprises a single forward and a single reverse FASTQ file containing 7,632,013 x 150bp long reads totalling 1,144,801,950bp each.

Dermacentor variabilis

The dataset titled 'Whole-genome sequencing of dog tick, *Dermacentor variabilis*' is an Illumina HiSeq 2500 paired-end whole body shotgun sequencing run on an American dog tick (*D. variabilis*) submitted to NCBI by Johns Hopkins University on the 31st of December 2017. The compressed 15.8Gb dataset is made up of a single forward and a single reverse FASTQ file containing 224,784,714 101bp long reads totalling 22,703,256,114bp each.

Tick Virome

The dataset titled 'Virome of ticks:Tick06' is an Illumina MiSeq paired-end shotgun sequencing run of a randomly pooled sample of 50 ticks from Liaoning, China, and was submitted by Jiangsu University on the 1st of December 2020. The compressed 25.4Mb dataset comprises a single forward and a single reverse FASTQ file containing 86,003 x 250bp long reads totalling 19,165,358bp each.

2.4 - Quality Control and Filtering

The raw paired sequences of each lane were assessed for sequence quality, base quality, GC content, N content, duplication levels and over-represented sequences using FastQC (Andrews, 2010) on the Galaxy Bioinformatics platform (Blankenberg et al., 2010). The output was then aggregated using MutliQC (Ewels et al., 2016) to provide a succinct integrative report capable of comparing the data. Once initial read quality was confirmed, the datasets obtained from NCBI SRA and AGRF were then further analysed using the all-in-one read processor fastp for quality profiling, adapter trimming, read filtering and base correction, using default trimming parameters (Chen et al., 2018). Reads with a phred quality score lower than 15 and a read length shorter than 15bp were removed from downstream analysis. Using USEARCH (filter_phix), the datasets were analysed for any potential contaminated PhiX sequencing controls that were to be removed during sequencing (Edgar, 2010).

2.5 - Read Isolation

A first stage draft genome was produced using the program MEGAHIT v1.2.9 (Li et al., 2015) on each of the datasets from AGRF and NCBI SRA used for this project, including *D. variabilis* (SRR5317834), *R. appendiculatus* (SRR11263445), *R. linnaei* (SRR13348027), *H. longicornis* (SRR13164901) and CTL-1 Lane 1 (AGRF) using the quality-controlled trimmed reads generated by fastp. The primary output file from MEGAHIT designated 'final.contigs.fna' was used for downstream analysis. MEGAHIT was run using the default parameters. The contigs generated from MEGAHIT were then quality assessed using Samtools for initial statistics and then metaQUAST to provide an assessment of assembly quality (Mikheenko et al., 2016).

All shotgun sequencing data were analysed using the read alignment software Bowtie2 and BWA-MEM algorithm to five reference genomes. The reference genomes were obtained from NCBI and included: *I. scapularis* (ID: 523)*, I. ricinus* (ID: 16267)*, R. microplus* (ID: 2797)*, Dermacentor silvarum* (ID: 35355)*, H. longicornis* (ID: 69202), *E. coli* (ID: 167) *and H. sapiens* (ID: 51). Bowtie2 was run using the 'very sensitive local' pre-set on the trimmed, paired-ended shotgun datasets from AGRF and NCBI. BWA-MEM was run to compare results with Bowtie2 alignments and used the default local alignment mapping parameters (Langmead & Salzberg, 2012; Li, 2013). The resulting Sequence Alignment Map (SAM) files from read alignments were converted to FASTA files using Samtools (view and fasta tools). The Bourne Again SHell (BASH) script that was used (mapped-reads_to_fasta.bash) can be found here: https://github.com/XWBarton/HonoursScripts.

The first five reference genomes were Ixodidae genomes that were tested against the shotgun datasets, to identify the best mapped reference genome for the recovery of tick genomic data. The two reference genomes, *Escherichia coli* and *Homo sapiens,* were used to align the tick datasets as a control. On one level, these reference genomes can show the quality of a sequencing run because if a primarily tick-based shotgun sequencing run was to be mapping at a high percentage to either *E. coli* or *H. sapiens,* (species that are both undesired in the sequencing run and have an unlikely

presence), then it can show that the run did not work as intended and the data would not be useful for downstream analysis. As well as this, there is some probability that human DNA contamination of a sequencing run can occur at the stage of sample collection or DNA extraction, so mapping to a human reference genome allows for a check of human contamination.

Four Kraken2 databases were constructed to compare how shotgun sequencing reads would be classified depending on what genomic content was added to each of the databases. The "Standard" database is the Kraken2 default database primarily used for microbial sequencing runs (Wood et al., 2019). The primary database used in this project was "tick-db" made up of RefSeq bacterial, viral and protozoan genomes, as well as of all Ixodidae genomes available on NCBI at the time of writing for tick content and a selection of marsupial genomes present on NCBI including the genera *Isoodon* (bandicoot), *Petrogale* (wallaby), *Notamacropus* (wallaby), *Vombatus* (wombat), *Sarcophilus* (Tasmanian devil) and *Trichosurus* (possum) genomes for potential host genomic content (see Table 2). The third Kraken2 database constructed "tick+human-db" was made of the entire tick-db contents and a human genome assembly to see how the results may be changed. Finally, "microbe-db" was constructed which contained only microbial content which may have been present in the tick sequencing data. This was used to compare how microbial content was classified between the different databases. The comparison of the content of the databases can be compared in Table 3.

The paired-end NovaSeq 6000 data and the SRA data was then processed by Kraken2 using the custom tick-focused database (tick-db), using the default k-mer parameter length of 35, minimum base quality of 0, and minimum hit group number of 2. The output text file report generated from Kraken2 was then visualised using KronaTools v2.7.1 using the taxa identification and read count column, producing an interactive radial HTML figure describing how many reads were classified to each taxon in the database.

The datasets were then also processed using the three other Kraken2 databases to find any differences in classification results.

Table 2. Genomes added to the custom tick-centric Kraken2 databases obtained from NCBI

Taxon	Organism	NCBI Genome ID	GenBank
Ixodidae	*Rhipicephalus annulatus*	93261	GCA_013436015.1 TxGen Rann
	Hyalomma asiaticum	92161	GCA_013339685.1 ASM1333968v1
	Haemaphysalis longicornis	69202	GCA_013339765.1
	Dermacentor silvarum	36355	GCA_013339745.1 ASM1333974v1
	Ixodes ricinus	16267	GCA_000973045.2
	Ixodes persulcatus	7207	GCA_013358835.1 BMI_IPER_1.0
	Rhipicephalus microplus	2797	GCF_013339725.1 ASM1333972v1
	Rhipicephalus sanguineus	2716	GCF_013339695.1 ASM1333969v1
	Ixodes scapularis	523	GCF_002892825.2 ISE6_asm2.2
Marsupialia	*Isoodon obesulus*	98179	GCA_904813085.1 ERS5079773
	Petrogale xanthopus	98174	GCA_904812355.1 ERS5079696
	Isoodon auratus	98147	GCA_904811015.1 ERS5079592
	Notamacropus eugenii	233	GCA_000004035.1 Meug_1.1
	Petrogale brachyotis	98177	GCA_904812665.1 ERS5079734
	Isoodon macrourus	7109	GCA_904811065.1 ERS5079590
	Trichosurus arnhemensis	98156	GCA_904811895.1 ERS5079657
	Vombatus ursinus	7180	GCA_900497805.2
	Sarcophilus harrisii	3066	GCF_902635505.1 mSarHar1.11

Table 3. Comparing Content and Size of Constructed Kraken2 Databases

Kraken2 Database	Contents	Size (Gb)
Standard (k-mer 35)	RefSeq for the bacterial, archaeal, and viral domains, along with the human genome and a collection of known vectors (UniVec_Core)	182
tick-db (k-mer 35)	RefSeq for bacterial, viral and protozoan domains. All Ixodidae tick genomes from NCBI and a selection of marsupial genomes obtained from NCBI (see Table 2.)	204
tick+human-db (k-mer 35)	All of tick-db plus *Homo sapiens* (assembly GRCh38.p13)	207
microbe-db (k-mer 35)	RefSeq for the bacterial, protozoal, and viral domains	165

2.6 - Compute Resources

The computing power for this project was made available through a Pawsey Supercomputing Centre internship for 10-week project to develop skills utilising supercomputational biology. This project used the Nimbus Cloud Computing Service with the 16 core, 64 GB RAM (Random Access Memory) flavour running the Ubuntu / UNIX operating system.

Small BASH scripts were created to ease data analysis and visualisation. Kronabuild was created to aid in the generating of Krona charts so that they could easily be converted from Kraken2 reports in

order to quickly see, in an effective manner, what organisms were classified in a dataset. Nicekrak was developed as an instant and easy way to see precisely how much tick was classified using only the Kraken2 output report. A BASH script to convert a FASTQ file to a FASTA file was created using the inbuilt UNIX program sed, as well as a BASH script to convert reads aligned to a reference genome in a SAM file (from Bowtie2 or BWA-MEM) into a FASTA format. Finally, tick builder was created using the Kraken2 database build program to easily create the custom tick focused Kraken2 database used primarily in this project. This script is beneficial as the database exceeds 200 GBs in size (see Table 3), making it impractical for transferring to other systems. The scripts use significantly less storage by automatically setting up directories and downloading the required genomes. As new genomes become available, relevant species can be added to the script, improving the performance of the Kraken2 database. These scripts can be viewed at: https://github.com/XWBarton/HonoursScripts.

2.7 - Primer Selection

The program 'STR detection' (Fungtammasan et al., 2015) was used to identify potential microsatellite sequences in the filtered and un-filtered datasets. Primers were designed according to the following parameters: product size: 100-300; minimum primer size: 18 maximum primer size: 27; optimal primer size: 20; minimum primer melting temperature: 58°C; maximum primer melting temperature: 65°C; optimum primer melting temperature: 60°C. The output sequences from STR detection were imported into Geneious v 8.0.5 (https://www.geneious.com) to assess suitability for primer design. Flanking regions of microsatellites were assessed visually to identify if they would be a suitable location for primer binding. Reads that were not suitable were not used for primer design; these included flanking regions that were generic or repetitive. Primers were then designed for the suitable flanking regions using the Geneious primer designing tool. Primers were then ordered from Integrated DNA Technologies. In total, 19 microsatellite primer pairs were developed and ordered.

Krait is an open-source graphical user interface program that can be run on Windows, Mac or Linux to scan for microsatellites sequences and their flanking regions and designs primers for the flanking regions (Du et al., 2018). After the fastp trimming, MEGAHIT assembly, Bowtie2 alignment and Kraken2 classification of datasets, the output files of these programs were converted to the FASTA format and downloaded from the Nimbus Cloud Computer to a local machine running Windows 10. These sequences were then parsed by Krait, searching for all mono-nucleotide, di-nucleotide, tri-nucleotide, tetra-nucleotide, penta-nucleotide and hexa-nucleotide motifs. After this, the identified microsatellite sequences were filtered using the inbuilt table filter tool to identify microsatellite sequences that had a beginning flanking region of larger than 20 base pairs, a microsatellite repeat length of larger than 50 base pairs and a microsatellite motif length of greater than two base pairs (tri-nucleotide or above). Of these filtered microsatellite sequences, the Primer3 tool built into Krait was used to develop primer sets using the default parameters, these primer sets along with the associated Tm, motif, length and stability rating, were exported to Excel.

2.8 - Primer Testing

DNA Extraction

Previously extracted tick genomic DNA samples were used to test against the primers created by 'STR Detection'. Extractions were performed by Dr Alex Gofton on *I. holocyclus* ticks obtained from the Sydney area, Australia using the QIAGEN DNeasy Blood and Tissue Kit (QIAGEN, Germany) following the manufacturer's supplementary protocol "Purification of total DNA from insects using the DNeasy Blood & Tissue Kit" kit.

PCR Assays

First, an Ixodidae 12S barcoding assay was performed to confirm the presence of tick DNA in the extractions (*n* =6). This was performed using the primers T1B (5'- AAACTAGGATTAGATACCCT-3' and T2A (5'- AATGAGAGCGACGGGCGATGT-3'), which amplifies a ~370bp fragment of the 12S gene and identifies all Ixodidae (hard ticks) (Beati & Keirans, 2001). The PCR was performed in 25µL volumes

and consisted of 18.5µL of DNA free water, 2.5µL of 1X (1.5mM) KAPA Taq Buffer + dye (with MgCl2), 1µL of 1.0mM KAPA $MgCl_2$, 1µL of 0.4µM T1B forward primer, 1µL of 0.4µM T2A reverse primer, 0.25µL of 0.25mM dNTPs (Fisher Biotech), 0.1µL / 0.5 U KAPA Taq and 2µL of genomic DNA. The samples were screened using the AppliedBiosystems SimpliAmp PCR Thermocycler using firmware v1.3.4. Thermocycling conditions included an initial hold for 5 mins at 94°C followed by 5 cycles of denaturation at 94°C for 15 secs, annealing at 51°C for 30 secs and extension at 68°C for 30 secs, after this there was 25 cycles of denaturation at 94°C for 30 secs, annealing at 53°C for 30 secs and extension at 70°C for 1 min with a final extension of 70°C for 5 mins.

Developed microsatellite primers (from 'STR detection') were then tested using the *I. holocyclus* gDNA (*n* = 6). The PCRs were performed in 25µL volumes, which consisted of 16.9µL of DNA free water, 2.5µL of 1X (1.5mM) KAPA Taq Buffer + dye (with MgCl2), 1µL of 1.0mM KAPA $MgCl_2$, 1µL of 0.4µM Forward microsatellite primer, 1µL of 0.4µM Reverse microsatellite primer, 0.5µL of 0.25mM dNTPs (Fisher Biotech), 0.1µL / 0.5 U of KAPA Taq and 2µL of genomic DNA. The samples were screened using the AppliedBiosystems SimpliAmp PCR Thermocycler using firmware v1.3.4. Thermocycling conditions included an initial hold step at 95°C for 5 mins, then followed by 50 cycles of denaturation at 95°C for 30 secs, annealing at 55°C for 30 secs and extension at 72°C for 1 min with a final extension of 72°C for 5 mins. After this, the same set of developed microsatellite primers were tested using the same protocol except for the addition of an extra 2µL of 1.0mM KAPA $MgCl_2$ and a reduction in 2µL of DNA free water.

Gel Electrophoresis and Sanger Sequencing

Microsatellite PCR products were then run on an agarose gel electrophoresis. 1.0 gram of agarose gel powder was dissolved into 100mL of heated 1X TAE Buffer to produce a 1% (w/v) gel solution to which then 1.5 ml Axygen SYBR Safe DNA stain was added, the whole mixture was then poured into a gel mould. 20µL of PCR product was added to each agarose gel well along with 5µL of Axygen 100bp and 1kbp DNA ladders. The gel was then run at 80V for 50 mins. The gel was then viewed

under UV light, and positive bands were excised using a scalpel blade. Positive bands identified on the gel were purified (Yang et al., 2013) and sent to AGRF for Sanger sequencing. Sequences were then to be analysed using Geneious to determine if microsatellite regions were amplified.

Figure 1. Read Isolation and Primer Development Bioinformatics Workflow

3 - Results

3.1 - Quality Control and Filtering

***Ixodes holocyclus* AGRF data – Lane 1**

The forward and reverse FASTQ Lane 1 shotgun sequence reads were assessed and filtered using fastp. The duplication rate was assessed at 7.2%. The FASTQ files contained 20,203,828 total reads and 5,050,957,000 total bases before quality filtering. After quality filtering, the FASTQ files contained 19,956,978 reads and 4,524,442,000 bases with 10.42% of bases being removed. Of the reads that did not pass the quality filter, 234,730 were due to low quality (a quality score lower than 20), 1,596 reads were removed for being too short (less than 15bp) and 10,524 were due to too many Ns. 6,052,708 reads in the dataset also underwent adapter trimming. Filtering for PhiX showed 12 hits, being 6.013×10^{-7}% of the sequence data. The Q30 score after filtering was 90.8%.

Rhipicephalus linnaei

The forward and reverse FASTQ *R. linnaei* NCBI SRA files were assessed and filtered using fastp. The duplication rate was assessed at 18.0599%. The FASTQ files contained 12,619,829 total reads and 1,892,974,350 total bases each, before quality filtering. After quality filtering the FASTQ files contained 12,449,000 reads and 1,857,375,983 bases each with 1.88% of bases being removed. In total, 24,898,000 reads passed quality filter (Table 4). Of the reads that did not pass the quality filter, 341,336 were due to low quality (a quality score lower than 20) and 322 were due to too many Ns. 357,120 reads in the dataset also underwent adapter trimming. Filtering for PhiX showed no hits. The Q30 score after filtering was 92.9%.

Rhipicephalus appendiculatus

The forward and reverse reads from the FASTQ dataset of *R. appendiculatus* were assessed and filtered using fastp. The duplication rate was assessed at 14.4363%. The FASTQ files contained 7,632,013 total reads and 1,144,801,950 total bases each, before quality filtering. After quality

filtering the FASTQ files contained 7,448,218 reads and 1,106,381,210 bases each, with 3.36% of the bases being removed. In total, 14,896,436 reads passed quality filter (Table 4). Of the reads that did not pass quality filter, 367,484 were due to too low quality (a quality score lower than 20) and 106 were removed due to too many Ns. 440,088 reads in the dataset also underwent adapter trimming. Filtering for PhiX showed no hits. The Q30 score after filtering was 91.9%.

Dermacentor variabalis

The forward and reverse reads from the FASTQ dataset of *D. variabalis* were assessed and filtered using fastp. The duplication rate was assessed at 0.0550251%. The FASTQ files contained 224,784,714 total reads and 22,703,256,114 total bases each, before quality filtering. After quality filtering the FASTQ files contained 215,002,510 reads and 21,527,061,491 bases each, with 5.18% of the bases being removed. In total, 430,005,020 passed the quality filter (Table 4). Of the reads that did not pass quality filter, 19,182,852 were due to low quality (a quality score lower than 20) and 381,556 were due to too many Ns. 12,668,490 reads in the dataset also underwent adapter trimming. Filtering for PhiX showed 33,680 hits being 7.83×10^{-3}% of the sequence data. The Q30 score after filtering was 94.7%.

Tick Virome

The forward and reverse reads from the FASTQ dataset of the Tick Virome were assessed and filtered using fastp. The duplication rate was assessed at 2.02956%. The FASTQ files contained 86,003 total reads and 19,081,453 total bases each, before quality filtering. After quality filtering the FASTQ files contained 79,835 reads and 17,139,158 bases each, with 10.18% of bases being removed. In total, 159,670 passed quality filter (Table 4). Of the reads that did not pass quality filter, 12,322 were due to low quality (a quality score lower than 20) and 14 were due to too many Ns. 10,738 reads in the dataset also underwent adapter trimming. Filtering for PhiX showed no hits. The Q30 score after filtering was 80.8%.

Table 4. Read Count Before and After Quality Control used Datasets

Lane	Total Reads Before Filtering (F + R)	Total Reads After Filtering (F + R)
Lane 1	10,022,726 + 10,022,726	9,897,893 + 9,897,893
R. linnaei	12,619,829 + 12,619,829	12,449,000 + 12,449,000
R. appendiculatus	7,632,013 + 7,632,013	7,448,218 + 7,448,218
D. variabalis	224,784,714 + 224,784,714	215,002,510 + 215,002,510
Tick Virome	86,003 + 86,003	79,835 + 79,835

3.2 - MEGAHIT Genome Assembly

MEGAHIT sequencing results were generated for optimised downstream analysis. Samtools produced the initial statistics, showing differences in assembly between the datasets, showing the varying numbers of contigs generated, sizes of contigs and reads that were assembled. A summary of the largest contig size, number of contigs generated, number of bases generated and N50 score of each dataset can be seen in Table 10. Further statistics were produced by metaQUAST (Mikheenko et al., 2016).

Ixodes holocyclus AGRF data – Lane 1

Lane 1 sequencing data from *I. holocyclus* DNA (AGRF) produced 1,475,232 contigs, with the largest contig generated being 22,457 bp in length, making up a total of 938,932,621 bp. The N50 statistic was 710 bp. Assembly quality of the contigs generated from Lane 1 was also assessed using metaQUAST. The report from this assessment confirms the contig statistics from Samtools and provides the distribution of contig size (*see* Table 5). In Lane 1, 13.17% of the contigs were greater than 999 bp in length and only six being 10,000 bp or greater.

Table 5. Distribution of Contig Length of Lane 1 MEGAHIT assembly

Contig Length	Number of Contigs	Number of bp
contigs (>= 0 bp)	1,475,232	938,932,621
contigs (>= 1000 bp)	194,325	281,365,102
contigs (>= 5000 bp)	224	1,366,886
contigs (>= 10000 bp)	6	84,714
contigs (>= 25000 bp)	0	0
contigs (>= 50000 bp)	0	0

Rhipicephalus linnaei

The *R. linnaei* MEGAHIT contigs were assessed using samtools, which shows that 668,009 contigs were generated making up a total of 349,006,875 bp, with the largest contig being 75,962 bp in size. The N50 statistic was 538 bp. The contigs generated were also assessed using metaQUAST, the report from this assessment confirms the contig statistics from samtools and provides the distribution of read size (*see* Table 6), with 3.82% of the reads being greater than 999 bp in length and 58 being 10000 bp or greater.

Table 6. Distribution of Contig Length of R. linnaei MEGAHIT assembly

Contig Length	Number of Contigs	Number of bp
contigs (>= 0 bp)	668,009	349,006,875
contigs (>= 1000 bp)	25,518	38,393,659
contigs (>= 5000 bp)	424	3,194,562
contigs (>= 10000 bp)	58	872,423
contigs (>= 25000 bp)	2	108,640
contigs (>= 50000 bp)	1	75,962

Rhipicephalus appendiculatus

The *R. appendiculatus* MEGAHIT contigs were assessed using samtools, which shows that 282,982 contigs were generated making up a total of 140,551,078 bp, with the largest contig being 446,284 bp long. The N50 statistic of these contigs is 502 bp. The contigs generated were also assessed using metaQUAST, the report from this assessment confirms the contig statistics from samtools and provides the distribution of read size (*see* Table 7), with 2.93% of the reads being greater than 999 bp in length and ten being 10,000 bp or greater.

Table 7. Distribution of Contig Length of R. appendiculatus MEGAHIT assembly

Contig Length	Number of Contigs	Number of bp
contigs (>= 0 bp)	282,982	140,551,078
contigs (>= 1000 bp)	8,298	13,971,103
contigs (>= 5000 bp)	88	1,940,019
contigs (>= 10000 bp)	10	1,473,387
contigs (>= 25000 bp)	6	1,425,951
contigs (>= 50000 bp)	5	1,396,347

Dermacentor variabilis

The *D. variabilis* MEGAHIT contigs were assessed using samtools, which shows that 2,493,630 contigs were generated making up a total of 2,541,045,704 bp, with the largest contig being 289,192 bp long. The N50 statistic of these contigs is 2,184 bp. The contigs generated were also assessed using metaQUAST. The report from this assessment confirms the contig statistics from samtools and provides the distribution of read size (*see* Table 8), with 24.35% of the reads being greater than 999 bp in length and 15,097 being 10,000 bp or greater.

Table 8. Distribution of Contig Length of D. variabilis MEGAHIT assembly

Contig Length	Number of Contigs	Number of bp
contigs (>= 0 bp)	2,493,630	2,541,045,704
contigs (>= 1000 bp)	607,141	1,758,313,978
contigs (>= 5000 bp)	76,973	630,754,127
contigs (>= 10000 bp)	15,097	214,939,552
contigs (>= 25000 bp)	531	18,415,947
contigs (>= 50000 bp)	37	3,358,179

Tick Virome

The Tick Virome MEGAHIT contigs were assessed using samtools, which shows that 4,956 contigs were generated making up a total of 2,195,478 bp, with the largest contig being 3,068 bp in size. The N50 statistic was 438 bp. The contigs generated were also assessed using metaQUAST, the report from this assessment confirms the contig statistics from samtools and provides the distribution of read size (*see* Table 9), with 0.97% of the contigs being less than 999 bp in length and only 0 being 5,000 bp or greater.

Table 9. Distribution of Contig Length of Tick Virome MEGAHIT assembly

Contig Length	Number of Contigs	Number of bp
contigs (>= 0 bp)	4,956	2,195,478
contigs (>= 1000 bp)	48	65,946
contigs (>= 5000 bp)	0	0
contigs (>= 10000 bp)	0	0
contigs (>= 25000 bp)	0	0
contigs (>= 50000 bp)	0	0

Table 10. Comparison of MEGAHIT Assembly Metrics Across Different Datasets Produced by Samtools

Dataset (SRA Accession Number)	Number of Contigs	Largest Contig (bp)	Total base pairs (bp)	N50 (bp)
Lane 1	1,475,232	22,457	938,932,621	710
R. linnaei	668,009	75,962	349,006,875	538
D. variabilis	2,493,630	289,192	2,541,045,704	2,184
R. appendiculatus	282,982	446,284	140,551,078	502
Tick Virome	4,956	3,068	2,1954,78	438

3.3 - Read Mapping

Unassembled trimmed data was mapped to seven different reference genomes using the programs Bowtie2 and BWA (MEM algorithm) (*see* Section 2.5). Table 11 and Table 12 surmise the differences in the total alignment percentage between the different datasets, read mapping programs and reference genomes. Figure 2 and Figure 3 display the total alignment percentage across different datasets and reference genomes for Bowtie2 and BWA (MEM) respectively, highlighting the discrepancy in total alignment percentage between the read alignment programs. These tables and figures can be found at the end of Section 3.3.

***Ixodes holocyclus* AGRF data – Lane 1**

In total, 19,795,786 unassembled quality-controlled reads in Lane 1 were mapped using Bowtie2 and BWA-MEM to the five reference genomes. When mapping using Bowtie2, 418,307 (2.11%) reads were aligned when mapping to *I. scapularis*, 297,317 (1.50%) reads were aligned when mapping to *I. ricinus*, 404,469 (2.04%) reads were aligned when mapping to *R. microplus*, 305,088 (1.54%) reads were aligned when mapping to *D. silvarum*, 433,663 (2.19%) reads were aligned when mapping to *H. longicornis*, 121 (0.00%) reads were aligned when mapping to *E. coli* and 108,939 (0.55%) reads were aligned when mapping to *H. sapiens*.

When mapping using BWA-MEM, 11,242,691 (56.72%) reads were aligned when mapping to *I. scapularis*, 7,919,083 (39.96%) reads were aligned when mapping to *I. ricinus*, 11,975,029 (60.32%) reads were aligned when mapping to *R. microplus*, 11,469,335 (57.88%) reads were aligned when mapping to *D. silvarum*, 12,143,018 (61.20%) reads were aligned when mapping to *H. longicornis,*

135 (0.00%) reads were aligned when mapping to *E. coli* and 5,367,119 (27.05%) reads were aligned when mapping to *H. sapiens*.

Rhipicephalus linnaei

In total, 24,898,000 unassembled quality-controlled reads for *R. linnaei* were mapped using Bowtie2 and BWA-MEM to the five reference genomes. When mapping using Bowtie2, 720,216 (2.89%) reads were aligned when mapping to *I. scapularis*, 546,181 (2.1%) reads were aligned when mapping to *I. ricinus*, 10414768 (41.83%) reads were aligned when mapping to *R. microplus*, 20.58 (20.58%) reads were aligned when mapping to *D. silvarum*, 2,044,902 (8.21%) reads were aligned when mapping to *H. longicornis*, 111 (0.0%) reads were aligned when mapping to *E. coli* and 334,250 (1.34%) reads were aligned when mapping to *H. sapiens*.

When mapping using BWA-MEM, 3,569,799 (14.32%) reads were aligned when mapping to *I. scapularis*, 2,440,284 (9.79%) reads were aligned when mapping to *I. ricinus*, 15,534,513 (58.49%) reads were aligned when mapping to *R. microplus*, 9,504,321 (37.68%) reads were aligned when mapping to *D. silvarum*, 5,830,970 (23.32%) reads were aligned when mapping to *H. longicornis,* 177 (0.00%) reads were aligned when mapping to *E. coli* and 3,436,281 (13.77%) reads were aligned when mapping to *H. sapiens*.

Rhipicephalus appendiculatus

In total, 14,896,436 unassembled quality-controlled reads for *R. appendiculatus* were mapped using Bowtie2 and BWA-MEM to the five reference genomes. When mapping using Bowtie2, 1,290,039 (8.66%) reads were aligned when mapping to *I. scapularis*, 1,091,913 (7.33%) reads were aligned when mapping to *I. ricinus*, 7,621,154 (51.16%) reads were aligned when mapping to *R. microplus*, 3,732,234 (25.05 %) reads were aligned when mapping to *D. silvarum*, 2,404,569 (16.14%) reads were aligned when mapping to *H. longicornis,* 755 (0.01%) reads were aligned when mapping to *E. coli* and 955,972 (6.42%) reads were aligned when mapping to *H. sapiens*.

When mapping using BWA-MEM, 3,316,592 (22.02%) reads were aligned when mapping to *I. scapularis*, 2,968,508 (19.74%) reads were aligned when mapping to *I. ricinus*, 11,332,571 (69.37%) reads were aligned when mapping to *R. microplus*, 6,930,973 (45.70%) reads were aligned when mapping to *D. silvarum*, 5,431,307 (36.04%) reads were aligned when mapping to *H. longicornis,* 977 (0.01%) reads were aligned when mapping to *E. coli* and 3,536,659 (23.53%) reads were aligned when mapping to *H. sapiens*.

Dermacentor variabilis

In total, 430,005,020 unassembled quality-controlled reads in *D. variabilis* were mapped using Bowtie2 and BWA-MEM to the five reference genomes. When mapping using Bowtie2, 7,623,731 (1.77%) reads were aligned when mapping to *I. scapularis*, 5,299,783 (1.23%) reads were aligned when mapping to *I. ricinus*, 32,930,694 (7.66%) reads were aligned when mapping to *R. microplus*, 162,255,189 (37.73%) reads were aligned when mapping to *D. silvarum*, 21,987,453 (5.11%) reads were aligned when mapping to *H. longicornis,* 41,613 (0.01%) reads were aligned when mapping to *E. coli* and 4,671,337 (1.09%) reads were aligned when mapping to *H. sapiens*.

When mapping using BWA-MEM, 56,267,677 (13.08%) reads were aligned when mapping to *I. scapularis*, 34,700,694 (8.07%) reads were aligned when mapping to *I. ricinus*, 49,756,966 (21.30%) reads were aligned when mapping to *R. microplus*, 234,524,116 (54.12%) reads were aligned when mapping to *D. silvarum*, 79,580,710 (18.49%) reads were aligned when mapping to *H. longicornis,* 223,174 (0.05%) reads were aligned when mapping to *E. coli* and 62,330,836 (14.48%) reads were aligned when mapping to *H. sapiens*.

Tick Virome

In total, 159,670 unassembled quality-controlled reads in the tick virome were mapped using Bowtie2 and BWA-MEM to the five reference genomes. When mapping using bowtie2, 29,182 (18.28 %) reads were aligned when mapping to *I. scapularis*, 21,007 (13.16%) reads were aligned

when mapping to *I. ricinus*, 36,826 (23.06%) reads were aligned when mapping to *R. microplus*, 36,322 (22.75%) reads were aligned when mapping to *D. silvarum*, 139,800 (87.56%) reads were aligned when mapping to *H. longicornis,* 3,590 (2.25%) reads were aligned when mapping to *E. coli* and 23,750 (14.87%) reads were aligned when mapping to *H. sapiens*.

When mapping using BWA-MEM, 122,759 (72.70%) reads were aligned when mapping to *I. scapularis*, 90,307 (54.82%) reads were aligned when mapping to *I. ricinus*, 132,767 (78.38%) reads were aligned when mapping to *R. microplus*, 133,019 (78.23%) reads were aligned when mapping to *D. silvarum*, 186,502 (96.95%) reads were aligned when mapping to *H. longicornis,* 5,829 (3.64%) reads were aligned when mapping to *E. coli* and 126,321 (75.91%) reads were aligned when mapping to *H. sapiens*.

Table 11. Heat map of the percentage of quality-controlled reads that are mapped to each reference genome relative to the percentages in each row, using Bowtie2 and BWA-MEM

Bowtie2	***I. scapularis***	***I. ricinus***	***R. microplus***	***D. silvarum***	***E. coli***	***H. sapiens***	***H. longicornis***
Lane 1	2.11	1.5	2.04	1.54	0	0.55	2.19
R. linnaei	2.89	2.1	41.83	20.58	6.42	1.34	8.21
R. appendiculatus	8.66	7.33	51.16	25.05	0.01	6.42	16.14
D. variabalis	1.77	1.23	7.66	37.73	0.01	1.09	5.11
Tick Virome	18.28	13.16	23.06	22.75	2.25	14.87	87.56
BWA-MEM	***I. scapularis***	***I. ricinus***	***R. microplus***	***D. silvarum***	***E. coli***	***H. sapiens***	***H. longicornis***
Lane 1	56.72	39.96	60.32	57.88	0	27.05	61.2
R. linnaei	14.32	9.79	58.49	37.68	0	13.77	23.32
R. appendiculatus	22.02	19.74	69.37	45.7	0.01	23.53	36.04
D. variabalis	13.08	8.07	21.3	54.12	0.05	14.48	18.49
Tick Virome	72.7	54.82	78.38	78.23	3.64	75.91	96.95

Colour Legend: Highest Percentage Mapped | Lowest Percentage Mapped

Table 12. Comparison of Reads Aligned of Each Dataset to each Reference Genome using bowtie2 and BWA-MEM

Dataset	***I. scapularis*** (%)		***I. ricinus*** (%)		***R. microplus*** (%)		***D. silvarum*** (%)		***E. coli*** (%)		***H. sapiens*** (%)		***H. longicornis*** *(%)*	
	bowtie2	*BWA*	*bowtie2*	*BWA*	*bowtie2*	*BWA*	*bowtie2*	*BWA*	*bowtie2*	*BWA*	*bowtie2*	*BWA*	*bowtie2*	*BWA*
Lane 1	2.11	56.72	1.50	39.96	2.04	60.32	1.54	57.88	0.00	0	0.55	27.05	2.19	61.20
R. linnaei	2.89	14.32	2.1	9.79	41.83	58.49	20.58	37.68	6.42	0.00	1.34	13.77	8.21	23.32
R. appendiculatus	8.66	22.02	7.33	19.74	51.16	69.37	25.05	45.70	0.01	0.01	6.42	23.53	16.14	36.04
D. variabalis	1.77	13.08	1.23	8.07	7.66	21.30	37.73	54.12	0.01	0.05	1.09	14.48	5.11	18.49
Tick Virome	18.28	72.70	13.16	54.82	23.06	78.38	22.75	78.23	2.25	3.64	14.87	75.91	87.56	96.95

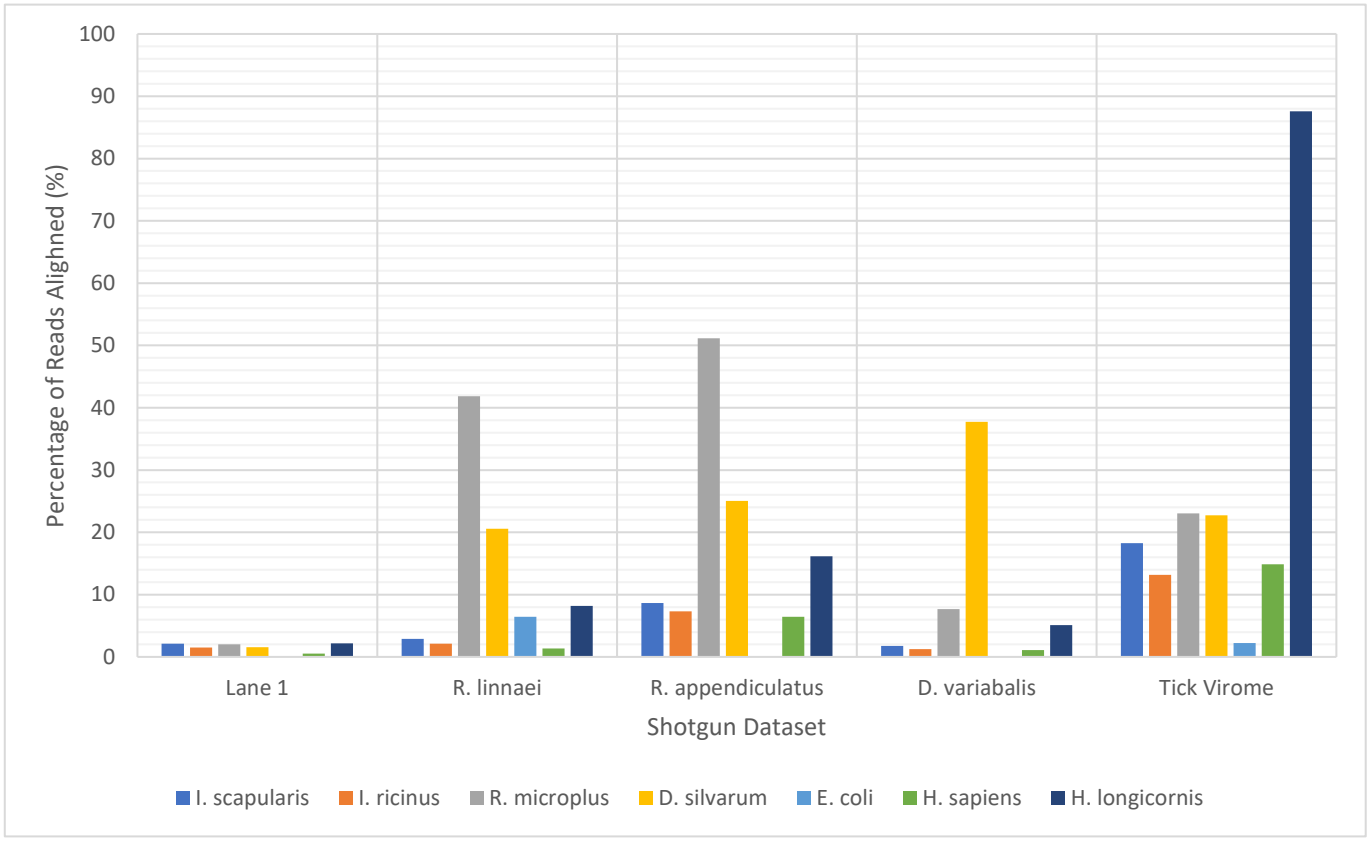

Figure 3. Percentage of Unassembled Reads Aligned from Each Dataset to Seven Different Reference Genomes using Bowtie2

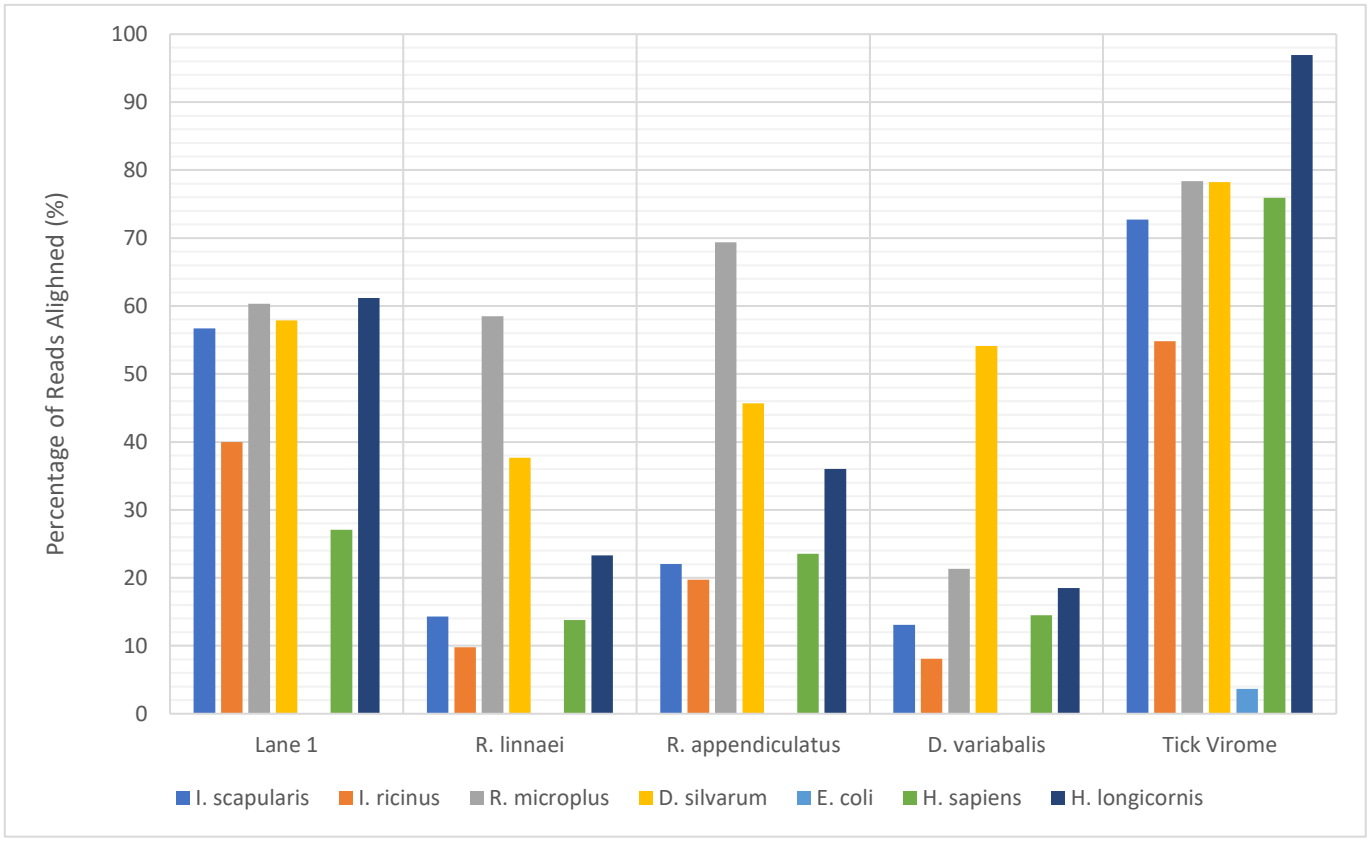

Figure 2. Percentage of Unassembled Reads Aligned from Each Dataset to Seven Different Reference Genomes using BWA-MEM

3.4 - Read Classification

Four initial databases were constructed for a comparison of how reads will be classified. The four databases were the Standard Kraken2 database, a custom microbial database (microbe-db), a tick focused database (tick-db) and a tick focused database with the addition of the human genome (tick+humad-db). All the SRA datasets were tested against each of the Kraken2 databases with varying result (*see* Section 2.3). The highest classification results were assessed using the BASH script 'nicekrak', which displays Ixodidae composition and percentage of sequences classified to each genus in descending order. Interactive HTML Krona charts (Figure 4.) for datasets classified using tick-db can be found at: https://xwbarton.github.io/KronaView/.

***Ixodes holocyclus* AGRF data – Lane 1**

Unassembled Lane 1 sequence reads were analysed using the four constructed Kraken2 databases. The top classification results of the Standard database were 1.79% classified to the genus *Homo*, 0.06% to the genus *Staphylococcus* and 0.03% to the genus *Streptomyces*. The tick-db was able to classify 4.34% (429,788 reads) of the sequences as Ixodidae (Figure 4.), with the top classification results being 1.56% *Ixodes*, 1.34% *Rhipicephalus* and 0.94% *Boophilus* (now considered *Rhipicephalus*). The tick+human-db was able to classify 4.33% (428,740 reads) of the sequences as Ixodidae, a -0.01% change from tick-db. The top classification results of tick+human-db were 1.55% *Ixodes*, 1.33% *Rhipicephalus* and 0.93% *Boophilus* (*Rhipicephalus*). The microbe-db top classification results were 0.06% *Staphylococcus*, 0.04% *Plasmodium* and 0.03% *Streptomyces*, which showed similar microbial content (with the exception of *Plasmodium* in microbe-db) to the Standard database.

The Lane 1 contigs assembled by MEGAHIT were also analysed using the four constructed Kraken2 databases. The top classification results for the Standard database were 2.24% *Homo*, 0.02% *Streptomyces* and 0.02% *Staphylococcus*. The tick-db was able to classify 5.26% of the sequences reads as Ixodidae, with the top classification results being 1.90% *Rhipicephalus*, 1.74% *Ixodes* and

1.39% *Boophilus* (*Rhipicephalus*). The tick+human-db was able to classify 5.27% of the sequence reads as Ixodidae, a +0.01% change from tick-db. The top classification results of tick+human-db were 1.90% *Rhipicephalus*, 1.74% *Ixodes* and 1.39% *Boophilus* (*Rhipicephalus*). Finally, the microbe-db top classification results were 0.02% *Staphylococcus*, 0.02% *Streptomyces* and 0.01% *Leishmania*, which showed similar results compared to the microbial content of the Standard database.

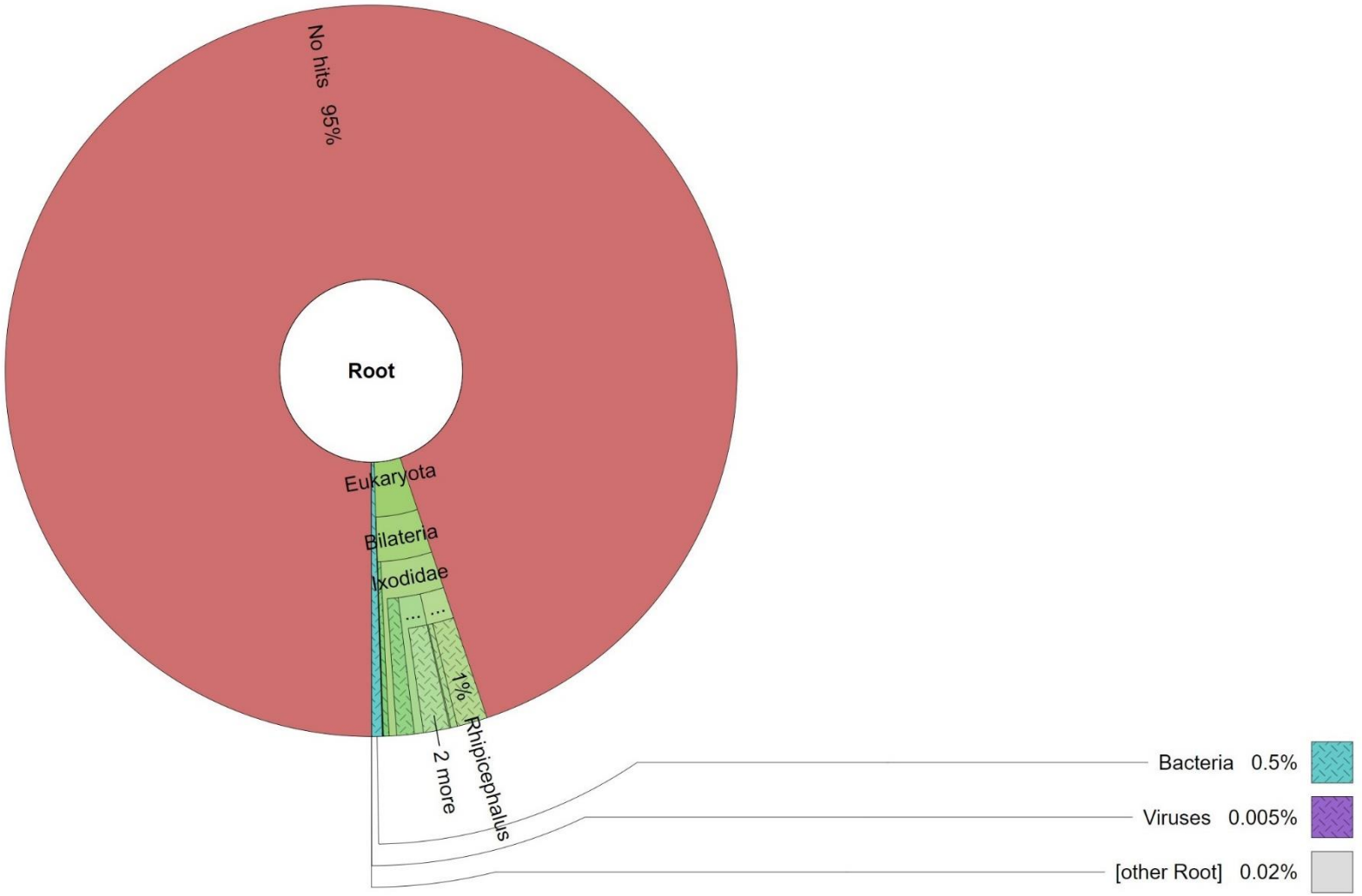

Figure 4 Metagenomics classification of Lane 1 using Kraken2 (tick-db). The figure has been produced by KronaTools v2.7.1 and displays rings that represent what percentage of reads were classified to a taxonomic group. As the rings move away from the centre of the figure, the classification becomes more specific to a taxonomic group. The red portion of the graph represents the percentage of reads that could not be classified, with the green segment representing Ixodidae classified in the sequence data. View the interactive chart with greater detail: https://xwbarton.github.io/KronaView/current-charts/1_tick-db_lane1.krona.html.

Rhipicephalus linnaei

Unassembled *R. linnaei* sequence reads and were analysed using the four constructed Kraken2 databases. The top classification results of the Standard database were 3.37% classified to the genus *Homo*, 0.30% to the genus *Coxiella* and 0.02% to the genus *Pseudomonas*. The tick-db was able to classify 97.47% (12,134,112 reads) of the sequences as Ixodidae, with the top classification results being 97.10% *Rhipicephalus*, 0.35% *Boophilus* (now considered *Rhipicephalus*) and 0.08% *Hyalomma*. The tick+human-db was able to classify 97.46 % of the sequences as Ixodidae, a -0.01% change from tick-db. The top classification results of tick+human-db were 97.10% *Rhipicephalus*, 0.35% *Boophilus* (*Rhipicephalus*) and 0.12% *Homo*. The microbe-db top classification results were 0.30% *Coxiella*, 0.02% *Plasmodium* and 0.02% *Pseudomonas*, which showed similar microbial content (with the exception of *Plasmodium* in microbe-db) to the Standard database.

The *R. linnaei* contigs assembled by MEGAHIT were also analysed using the four constructed Kraken2 databases. The top classification results for the Standard database were 3.32% *Homo*, 0.09% *Coxiella* and 0.02% *Streptomyces*. The tick-db was able to classify 98.56% of the sequences reads as Ixodidae (ticks), with the top classification results being 98.40% *Rhipicephalus*, 0.19% *Boophilus* (*Rhipicephalus*) and 0.06% *Dermacentor*. The tick+human-db was able to classify 98.56% of the sequence reads as Ixodidae, which was no change from tick-db. The top classification results of tick+human-db were 98.4% *Rhipicephalus*, 0.19% *Boophilus* (*Rhipicephalus*) and 0.06% *Dermacentor*. Finally, the microbe-db top classification results were 0.09% *Coxiella*, 0.01% *Streptomyces* and 0.01% *Pseudomonas*, which showed similar results compared to the microbial content of the Standard database.

Rhipicephalus appendiculatus

Unassembled *R. appendiculatus* sequence reads and were analysed using the four constructed Kraken2 databases. The top classification results of the Standard database were 10.71% classified to the genus *Homo*, 1.74% to the genus *Coxiella* and 0.19% to the genus *Clostridium*. The tick-db was

able to classify 61.88% of the sequences as Ixodidae, with the top classification results being 54.22% *Rhipicephalus*, 19.00% *Boophilus* (now considered *Rhipicephalus*) and 3.25% *Hyalomma*. The tick+human-db was able to classify 61.65% of the sequences as Ixodidae, a 0.23% change from tick-db. The top classification results of tick+human-db were 54.04% *Rhipicephalus*, 18.92% *Boophilus* (*Rhipicephalus*) and 3.23% *Hyalomma*. The microbe-db top classification results were 1.69% *Coxiella*, 0.11% *Babesia* and 0.09% *Clostridium*, which was very similar compared with the Standard database, with the exception of the protozoan parasite *Babesia*.

The *R. appendiculatus* contigs assembled by MEGAHIT were also analysed using the four constructed Kraken2 databases. The top classification results for the Standard database were 7.02% *Homo*, 0.16% *Clostridium* and 0.02% *Streptomyces*. The tick-db was able to classify 73.13% of the sequences reads as Ixodidae, with the top classification results being 68.32% *Rhipicephalus*, 22.56% *Boophilus* (*Rhipicephalus*) and 2.30% *Hyalomma*. The tick+human-db was able to classify 73.07% of the sequence reads as Ixodidae, a -0.06% change from tick-db. The top classification results of tick+human-db were 68.29% *Rhipicephalus*, 22.54% *Boophilus* (*Rhipicephalus*) and 2.29% *Hyalomma*. Finally, the microbe-db top classification results were 0.09% *Clostridium*, 0.08% *Babesia* and 0.03% *Leishmania*, which showed similar results compared to the microbial content of the Standard database with the exception of *Babesia*.

Dermacentor variabilis

Unassembled *D. variabilis* sequence reads and were analysed using the four constructed Kraken2 databases. The top classification results of the Standard database were 1.81% classified to the genus *Homo*, 0.24% to the genus *Arsenophonus* and 0.04% to the genus *Francisella*. The tick-db was able to classify 31.30% of the sequences as Ixodidae, with the top classification results being 25.28% *Dermacentor*, 3.04% *Rhipicephalus* and 1.40% *Boophilus* (now considered *Rhipicephalus*). The tick+human-db was able to classify 31.28% of the sequences as Ixodidae, a -0.02% change from tick-db. The top classification results of tick+human-db were 25.28% *Dermacentor*, 3.03% *Rhipicephalus*

and 1.40% *Boophilus* (*Rhipicephalus*). The microbe-db top classification results were 0.22% *Arsenophonus*, 0.05% *Xanthomonas* and 0.04% *Francisella*, which was very similar compared with the Standard database microbial content with the exception of *Xanthomonas.*

The *D. variabilis* contigs assembled by MEGAHIT were also analysed using the four constructed Kraken2 databases. The top classification results for the Standard database were 9.38% *Homo*, 0.04% *Streptomyces* and 0.02% *Pseudomonas*. The tick-db was able to classify 67.73% of the sequences reads as Ixodidae, with the top classification results being 59.86% *Dermacentor*, 4.55% *Rhipicephalus* and 2.08% *Boophilus* (*Rhipicephalus*). The tick+human-db was able to classify 67.69% of the sequence reads as Ixodidae, a -0.04% change from tick-db. The top classification results of tick+human-db were 59.86% *Dermacentor*, 4.53% *Rhipicephalus* and 2.06% *Boophilus* (*Rhipicephalus*). Finally, the microbe-db top classification results were 0.05% *Leishmania*, 0.05% *Plasmodium* and 0.04% *Streptomyces*, which showed similar results compared to the microbial content of the Standard database with the exception of *Leishmania* and *Plasmodium*.

Tick Virome

Unassembled Tick Virome sequence reads and were analysed using the four constructed Kraken2 databases. The top classification results of the Standard database were 11.65% classified to the genus *Homo*, 1.42% to the genus *Halomonas* and 0.47% to the genus *Escherichia*. The tick-db was able to classify 83.49% of the sequences as Ixodidae, with the top classification results being 76.01% *Haemaphysalis*, 1.54% *Ixodes* and 1.37% *Rhipicephalus*. The tick+human-db was able to classify 83.45% of the sequences as Ixodidae, a -0.04% change from tick-db. The top classification results of tick+human-db were 76.00% *Haemaphysalis*, 1.54% *Ixodes* and 1.36% *Rhipicephalus*. The microbe-db top classification results were 2.40% *Plasmodium*, 1.26% *Halomonas* and 0.67% *Cryptomonas*, which was somewhat different to the microbial content generated from the Standard database.

The Tick Virome contigs assembled by MEGAHIT were also analysed using the four constructed Kraken2 databases. The top classification results for the Standard database were 4.76% *Homo*,

0.44% *Halomonas* and 0.22% *Escherichia*. The tick-db was able to classify 91.75% of the sequences reads as Ixodidae, with the top classification results being 89.71% *Haemaphysalis*, 0.67% *Rhipicephalus* and 0.67% *Ixodes*. The tick+human-db was able to classify 91.73% of the sequence reads as Ixodidae, a -0.02% change from tick-db. The top classification results of tick+human-db were 89.69% *Haemaphysalis*, 0.67% *Rhipicephalus* and 0.67% *Ixodes*. Finally, the microbe-db top classification results were 0.91% *Plasmodium*, 0.56% *Halomonas* and 0.30% *Toxoplasma*, which showed similar results compared to the microbial content of the Standard database with the exception of *Plasmodium* and *Toxoplasma*.

With the exception of Lane 1, there was an overall increase in the percentage of reads classified as Ixodida using MEGAHIT assembled contigs over unassembled reads. The most significant change was observed between the *D. variabilis* dataset, where 36.43% more reads were classified as Ixodidae, and the smallest change was from the *R. linnaei* dataset, which had a 1.3% improvement when classifying Ixodidae. Of all of these increases, the top classification result remained the same, with one outlier being the Lane 1 dataset. The Lane 1 dataset was the only one that had a change in the top Kraken2 (tick-db) classification result changing from *Ixodes* (1.56%) to *Rhipicephalus* (1.90%) Measuring the number of bases between MEGAHIT assembled and trimmed, unfiltered sequence reads, there is an overall trend that the total number of bases classified using MEGAHIT assembled data is lower than that of trimmed unfiltered sequences. The number of bases classified by Kraken2 (tick-db) of the MEGAHIT assembled data is between 98.04% and 68.13% less bases than the trimmed unassembled sequences classified by Kraken2 (tick-db) (Table 13.).

Table 13. The Change in the Number of Bases Classified between Trimmed Unfiltered Datasets and MEGAHIT Assembled Datasets

Data	Unassembled (bases)	Assembled (bases)	Percentage Change (%)
Lane 1	200530342	63908633	-68.13
R. linnaei	3624714932	86215085	-97.62
R. appendiculatus	1375459862	27004413	-98.04
D. variabilis	13581706866	2170053889	-84.02
Tick Virome	29255044	2015818	-93.11

Table 14. Top Three Classification Results of Unassembled (trimmed, unfiltered) Reads using the Four Constructed Kraken2 Databases

Dataset	1st Highest Classification Result (%)	2nd Highest Classification Result (%)	3rd Highest Classification Result (%)
Lane 1 – Standard Database	*Homo* (1.79)	*Staphylococcus* (0.06)	*Streptomyces* (0.03)
Lane 1 – tick-db	*Ixodes* (1.56)	*Rhipicephalus* (1.34)	*Boophilus* (0.94)
Lane 1 – tick+human-db	*Ixodes* (1.55)	*Rhipicephalus* (1.33)	*Boophilus* (0.93)
Lane 1 – microbe-db	*Staphylococcus* (0.06)	*Plasmodium* (0.04)	*Streptomyces* (0.03)
R. linnaei – Standard Database	*Homo* (3.37)	*Coxiella* (0.30)	*Pseudomonas* (0.02)
R. linnaei – tick-db	*Rhipicephalus* (97.10)	*Boophilus* (0.35)	*Hyalomma* (0.08)
R. linnaei – tick+human-db	*Rhipicephalus* (97.10)	*Boophilus* (0.35)	*Homo* (0.12)
R. linnaei – microbe-db	*Coxiella* (0.30)	*Plasmodium* (0.02)	*Pseudomonas* (0.02)
R. appendiculatus – Standard Database	*Homo* (10.71)	*Coxiella* (1.74)	*Clostridium* (0.19)
R. appendiculatus – tick-db	*Rhipicephalus* (54.22)	*Boophilus* (19.00)	*Hyalomma* (3.25)
R. appendiculatus – tick+human-db	*Rhipicephalus* (54.04)	*Boophilus* (18.92)	*Hyalomma* (3.23)
R. appendiculatus – microbe-db	*Coxiella* (1.69)	*Babesia* (0.11)	*Clostridium* (0.09)
D. variabilis – Standard Database	*Homo* (1.81)	*Arsenophonus* (0.24)	*Francisella* (0.04)
D. variabilis – tick-db	*Dermacentor* (25.28)	*Rhipicephalus* (3.04)	*Boophilus* (1.40)
D. variabilis – tick+human-db	*Dermacentor* (25.28)	*Rhipicephalus* (3.03)	*Boophilus* (1.40)
D. variabilis – microbe-db	*Arsenophonus* (0.22)	*Xanthomonas* (0.05)	*Francisella* (0.04)
Tick Virome – Standard Database	*Homo* (11.65)	*Halomonas* (1.42)	*Escherichia* (0.47)
Tick Virome – tick-db	*Haemaphysalis* (76.01)	*Ixodes* (1.54)	*Rhipicephalus* (1.37)
Tick Virome – tick+human-db	*Haemaphysalis* (76.00)	*Ixodes* (1.54)	*Rhipicephalus* (1.36)
Tick Virome – microbe-db	*Plasmodium* (2.40)	*Halomonas* (1.26)	*Cryptomonas* (0.67)

Table 15. Top Three Classification Results of MEGAHIT Assembled Reads using the Four Constructed Kraken2 Databases

Dataset	1st Highest Classification Result (%)	2nd Highest Classification Result (%)	3rd Highest Classification Result (%)
Lane 1 – Standard Database	*Homo* (2.23)	*Streptomyces* (0.02)	*Staphylococcus* (0.02)
Lane 1 – tick-db	*Rhipicephalus* (1.90)	*Ixodes* (1.74)	*Boophilus* (1.39)
Lane 1 – tick+human-db	*Rhipicephalus* (1.90)	*Ixodes* (1.74)	*Boophilus* (1.39)
Lane 1 – microbe-db	*Staphylococcus* (0.02)	*Streptomyces* (0.02)	*Leishmania* (0.01)
R. linnaei – Standard Database	*Homo* (3.32)	*Coxiella* (0.09)	*Streptomyces* (0.02)
R. linnaei – tick-db	*Rhipicephalus* (98.40)	*Boophilus* (0.19)	*Dermacentor* (0.06)
R. linnaei – tick+human-db	*Rhipicephalus* (98.40)	*Boophilus* (0.19)	*Dermacentor* (0.06)
R. linnaei – microbe-db	*Coxiella* (0.09)	*Streptomyces* (0.01)	*Pseudomonas* (0.01)
R. appendiculatus – Standard Database	*Homo* (7.02)	*Clostridium* (0.16)	*Streptomyces* (0.02)
R. appendiculatus – tick-db	*Rhipicephalus* (68.32)	*Boophilus* (22.56)	*Hyalomma* (2.30)
R. appendiculatus – tick+human-db	*Rhipicephalus* (68.29)	*Boophilus* (22.54)	*Hyalomma* (2.29)
R. appendiculatus – microbe-db	*Clostridium* (0.09)	*Babesia* (0.08)	*Leishmania* (0.03)
D. variabilis – Standard Database	*Homo* (9.38)	*Streptomyces* (0.04)	*Pseudomonas* (0.02)
D. variabilis – tick-db	*Dermacentor* (59.86)	*Rhipicephalus* (4.55)	*Boophilus* (2.08)
D. variabilis – tick+human-db	*Dermacentor* (59.86)	*Rhipicephalus* (4.53)	*Boophilus* (2.06)
D. variabilis – microbe-db	*Leishmania* (0.05)	*Plasmodium* (0.05)	*Streptomyces* (0.04)
Tick Virome – Standard Database	*Homo* (4.76)	*Halomonas* (0.44)	*Escherichia* (0.22)
Tick Virome – tick-db	*Haemaphysalis* (89.71)	*Rhipicephalus* (0.67)	*Ixodes* (0.67)
Tick Virome – tick+human-db	*Haemaphysalis* (89.69)	*Rhipicephalus* (0.67)	*Ixodes* (0.67)
Tick Virome – microbe-db	*Plasmodium* (0.91)	*Halomonas* (0.56)	*Toxoplasma* (0.30)

3.5 - Microsatellite Detection

From adding the microsatellite primer sets developed from all datasets and categorising according to sequence isolation method, the following results were obtained: Unfiltered trimmed reads generated 578 primer sets, MEGAHIT assembled reads generated 2792 primer sets, Bowtie2 filtered reads generated 441 primer sets and Kraken2 filtered reads generated 617 primer sets (Figure 5). The HTML reports generated for these analyses, including additional statistics for microsatellite detection, can be found at: https://xwbarton.github.io/str-reports/str-reports.html. The microsatellite retrieval results for the datasets obtained from NCBI can be found in the Appendices, Section 6.1.

Ixodes holocyclus AGRF data – Lane 1

The program Krait v1.3.3 was able to detect a total of 473,379 STRs in the quality-controlled Lane 1 dataset; 153,155 mono-nucleotide, 111,442 di-nucleotide, 80,438 tri-nucleotide, 81,793 tetra-nucleotide, 42,494 penta-nucleotide and 4,057 hexa-nucleotide STR sequences. Out of the 473,379 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more, and a repeat length of 50 or more was a total of 2,450 STR sequences. Of the 2,450 filtered STR sequences, 70 had primer pairs successfully designed using Primer3 (integrated into Krait), according to the parameters described in the Materials and Methods (*see* Section 2.7). Parsing this data for total STRs took 3h 47min 8s.

After this, Krait was used the detect STRs from MEGAHIT assembled FASTA contigs from Lane 1. The program was able to detect 32,493 mono-nucleotide, 24,531 di-nucleotide, 18,736 tri-nucleotide, 18,226 tetra-nucleotide, 3,163 penta-nucleotide and 869 hexa-nucleotide STR sequences. Out of the 98,018 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 305 STR sequences. Of the 305 filtered STR

sequences, 190 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 14min 50s.

Krait was then used the detect STRs in the best-mapped data from bowtie2, which for Lane 1 was reads mapped to the reference genome *H. longicornis*. The SAM to FASTA converted reads were input into Krait. The program was able to detect 8,684 mono-nucleotide, 39,111 di-nucleotide, 21,999 tri-nucleotide, 21,394 tetra-nucleotide, 32,219 penta-nucleotide and 1,407 hexa-nucleotide STR sequences. Out of the 124,814 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 2,289 STR sequences. Of the 2,289 filtered STR sequences, 64 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 3min 8s.

Finally, reads classified as Ixodidae by Kraken2 using the custom tick-db database were analysed using Krait. The program was able to detect 4,187 mono-nucleotide, 5,539 di-nucleotide, 18,023 tri-nucleotide, 23,591 tetra-nucleotide, 30,508 penta-nucleotide and 453 hexa-nucleotide STR sequences. Out of the 82,301 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 1,636 STR sequences. Of the 1,636 filtered STR sequences, 35 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 7min 6s.

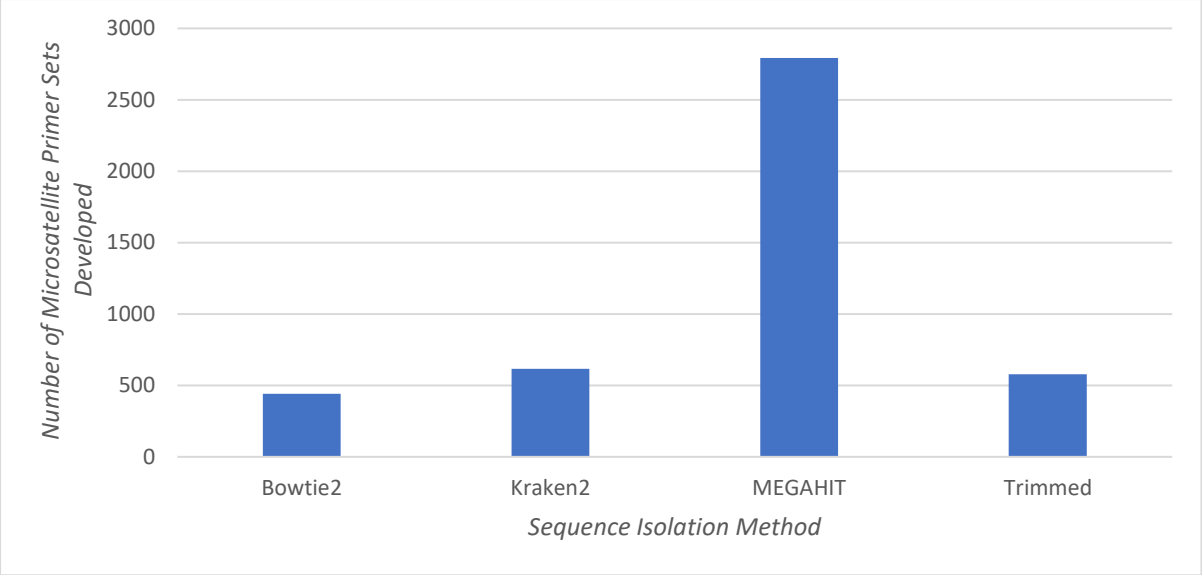

Figure 5. The Sum of the Microsatellite Primer Sets Developed using Krait v1.3.3 Across all Datasets from AGRF and NCBI

3.6 - PCR Analysis

Amplification was observed for all Ixodidae samples using the 12S barcoding assay, which was performed to confirm the presence and preservation of biobanked gDNA in *I. holocyclus* extractions. The no-template control displayed no amplification.

Of the 19 microsatellite primer sets developed from Lane 1 (*I. holocyclus*) using the program 'STR detection', no assays resulted in any amplification from any of the six *I. holocyclus* gDNA samples tested. This included experimentation with the standard KAPA Taq PCR and the KAPA Taq PCR with the addition of extra magnesium.

4 - Discussion

This project created a bioinformatics workflow to isolate specific taxonomic content from a shotgun NGS dataset from a mixed-species sample, for the purpose of designing microsatellite population genetic markers. The results obtained from this project have shown the many advantages over other more costly and time-intensive approaches. Overall, this workflow has the potential for dramatically reduced time required in wet-lab testing by narrowing the search for target sequences.

The original aims of this project were three-fold. Firstly, to organise reads from NGS mixed species samples into their appropriate taxonomic classification. Secondly, use in silico methods to identify potential microsatellite markers from the *I. holocyclus* NGS reads. Thirdly, test potential population genetics microsatellite markers on *I. holocyclus* DNA extractions to identify suitable markers. As this was a mixed species shotgun metagenomics dataset, there was additional complexity for the development of the bioinformatics workflow. Midway through the project, limitations were revealed with the AGRF generated dataset. Therefore, the aims of this project were modified to reflect this; firstly, confirm the quality and usability of the NGS data for potential downstream analysis; secondly, to organise reads from metagenomics dataset into their appropriate taxonomic classification; and finally, to use in silico methods to identify potential microsatellite markers from organised Ixodida shotgun NGS reads.

4.1 - Datasets and Quality Control

The quality control and filtering implemented into this pipeline were comparable to other bioinformatics workflows (Brinkmann et al., 2019; Mans et al., 2015; Tufts et al., 2020). These mainly involved checking the sample quality in features such as adapter and PhiX control contamination, sequencing yielded (number of bases generated), the average quality of bases within sequences and per base sequence contents. This standard quality assessment and control was performed using the program fastp which has been shown effective in other genomics (Masson et al., 2021; Veldsman et al., 2021) and metagenomic studies (Chen et al., 2021; Fan et al., 2021). As well as this, USEARCH

was used to filter for PhiX contamination as it has become common for highly contaminated sequences to be used for research in published literature (Mukherjee et al., 2015).

The dataset Lane 1 appeared to be satisfactory for analysis and retrieval of microsatellite markers, according to the statistics generated by fastp. Though during experimentation using genome assembly, read alignment and read classification the data was found to be unusable (*see* Section 4.7).

4.2 - Contig Assembly of Datasets (MEGAHIT)

The MEGAHIT assembler was used because of its relatively low power, conservative approach to assembling metagenomes and the recommendation for its accuracy in peer-reviewed comparisons of currently available assemblers (Forouzan et al., 2018). A conservative assembler is optimal for this workflow due to the lower number of false-positive assemblies produced. This creates a less developed assembly but is less critical as the sequences will be analysed, filtered and classified by taxon downstream in the workflow, and any misassembled microsatellite regions (performed by a less conservative assembler) would have the potential to cause unnecessary wet-lab testing.

The findings from this study are consistent with the assertion that an N50 score alone is unsatisfactory for the measurement of genome assembly quality (Bradnam et al., 2013). For instance, despite the differences in genome size and the size and number of contigs generated, the *R. appendiculatus* (containing 140,551,078 bp) and the Tick Virome datasets (containing 2,1954,78 bp) have similar N50 scores (of 502 bp and 428 bp). However, by looking at the distribution of contig sizes of the two datasets, it can be seen that *R. appendiculatus* has a better assembly (with 58 contigs larger than 10,000 bp) while the assembly of Tick Virome is relatively poor (with no contigs larger than 5,000 bp). It is therefore recommended to take other basic assembly statistics into account to provide a better assessment. Statistics that should be considered are number of contigs, size of largest/smallest contig, average length of contigs, total length of contigs and distribution of contig length (Yandell & Ence, 2012).

Although most datasets had an increase in the creation of microsatellite primers when the sequences were assembled by MEGAHIT, the dataset *D. variabilis* did not. This is likely due to the *D. variabilis* sequence reads being too short (101bp) to be efficiently assembled and therefore loosing many of the microsatellites from the original genome, providing evidence for the need for longer reads when generating microsatellites (*see* Section 4.7). However due to the significant increase in microsatellite primer sets able to be generated by the remaining datasets, it is recommended to try minimal assembly (computing power depending), if it is desired that microsatellite markers be developed from short-read next generation sequence data, to see if the microsatellite yield increases.

For the most part, performing an initial contig assembly appears to show a sizable increase in the number of microsatellite markers that can have primers designed. The sum total number of primers sets created using Krait and Primer3 from all assembled datasets was 2,792 primer sets. The dataset that produced the largest number of microsatellite primer sets using MEGAHIT assembled data was, *R. appendiculatus*, which produced 1,364 primer sets. This was 1,354 more primers than the quality-controlled unfiltered *R. appendiculatus* dataset. The smallest was the *D. variabilis* dataset that was not able to produce any primer sets. Microsatellite retrieval statistics for NCBI datasets can be found in the Appendix (Section 6.1). The significant increase from some of the datasets is understandable as having a larger sequence size (contigs) allows larger repeat regions to be detected by Krait if they were put together by the assembler. Alternatively, it means that larger flanking regions can be generated, making primer design easier.

4.3 - Read Alignment of Datasets (Bowtie2)

One aim of this project was to isolate specific taxonomic content from data, in order to better identify microsatellites. To do this a read alignment method was performed while testing the programs Bowtie2 and BWA-MEM. Seven reference genomes were downloaded from NCBI to

perform read alignment experimentation. These reference genomes allowed the retrieval of the most genomic content from each dataset, which were isolated from different genera.

Using *E. coli* as a negative control is not recommend in the future as *E. coli* DNA, though not explicitly reported present in the microbiome of the *I. holocyclus* DNA sample used in this study (S. Egan, 2017), is relatively abundant in the environment (Jang et al., 2017). Therefore, it has a chance of being present in the sample, for future experimentation such as this, a more biologically removed genome, perhaps from a plant would be more appropriate.

Bowite2 was run using the 'very-sensitive-local' pre-set to create the best chance for sequence reads to align to the reference genomes. This was done because preliminary testing showed bowtie2 was a more stringent read aligning program compared to others like BWA-MEM (Figures 2 and 3). This was optimal for this project as it meant that aligned reads could be confidently assumed to come from the reference genome. However, using a too stringent alignment algorithm would mean that potential microsatellite-containing reads may be removed from the data by not aligning. This is why the very-sensitive-local pre-set was used. The program BWA-MEM was first tested as it had been recommended based on its ability for optimised alignment of longer sequence reads (larger than 100 bp) compared to the previous iterations of the program. In addition to the software being a widely used and documented (Li, 2013; Thankaswamy-Kosalai et al., 2017). However, BWA-MEM was shown to be less stringent read alignment program for this project, where an over mapping of reads may result in non-tick microsatellite markers being developed. For example, datasets such as Lane 1 mapping to all tick reference genomes at a similarly high level, despite not having confidence that the sequencing run was successful. Due to the stringency results obtained, Bowtie2 was the program that was to be used for the workflow.

With the exception of the likely unsuccessful sequencing run in Lane 1, the variances of genomic data recovered from reads mapping appeared to correlate to the size of the dataset, with the larger datasets, mapping overall less to the reference genomes. *Dermacentor variabilis*, the largest dataset,

had the lowest level of aligned reads at 37.73% mapping the *D. silvarum* reference genome while Tick Virome, the smallest dataset had an alignment of 87.56% to the reference genome *H. longicornis*. Previous studies have evaluated the importance of sequencing depth in metagenomic samples, showing a reduction in sensitivity for microorganisms when a metagenomic sample is less deeply sequenced (Pereira-Marques et al., 2019). In the more deeply sequenced samples, all of the DNA present (including bacterial, protozoal and host) had an increased sensitivity to become amplified and be sequenced, possibly making a smaller proportion of the sequences tick. If this was the case, read classification using Kraken2 would be able to classify all microbiota in the sample. However, this did not occur with a similar proportion of sequences remaining unclassified.

Performing a sequence alignment step in a taxonomic isolation workflow appears to provide an adequate way of isolating specific genomic content. Prior studies have used read alignment techniques (Bowtie2 and BWA) to classify reads in tick whole-body shotgun data, containing both tick and associated microbes (Díaz-Sánchez et al., 2019; Tufts et al., 2020). However, as the field of metagenomics focuses on the microbiome, reads that were mapped to an Ixodida reference genome were removed for the downstream analysis of microbes, rather than kept for the downstream analysis of tick. Alternatively, in other studies, reads were mapped to bacterial genomes, and non-aligned reads (i.e. from tick) were removed (Carpi et al., 2011).

This method allows a significant amount of data to be recovered for downstream analysis once an appropriate reference genome is used. The minimum best mapped reference genome in this project was *D. variabilis* aligning to *D. silvarum*, which allowed the recovery of 37.73% of the dataset. Excluding dataset Lane 1, the average recovery rate for tick sequence data (using Bowtie2 and the best aligned reference genome) was 54.57%. This technique becomes increasingly inferior the more distantly related the reference genome is to the target organism. Therefore, it is useful to know, specifically, what the primary organism of the dataset is, to reduce significant amounts of time and computing power to find the best mapped reference genome. Together with this, the correct choice

in read alignment software must be made. If the software is too relaxed in its alignment, then many reads that are not of taxonomic interest may be used for downstream analysis, which is an inefficient use of time, developing and testing potential microsatellite markers. However, if the software is too stringent, microsatellite reads may be lost.

Across all datasets, using a read alignment technique produced a small decrease in the total microsatellite primer sets created compared to the unfiltered trimmed datasets. From of all datasets, the read alignment technique (using Bowtie2) created a total 441 microsatellite primer sets, compared to 578 from the unclassified trimmed datasets. The reduction in primers created is expected as sequences are removed from the dataset that are not classified as Ixodida from the best mapped to reference genome. These results indicate that using read mapping to create microsatellite marker primer sets for a target species will reduce the time required for wet-lab testing, by removing unnecessary microsatellite sequences.

4.4 - Read Classification of Datasets (Kraken2)

Kraken2 was chosen for this workflow for numerous reasons. Firstly, it is well cited (Kraken: 2466 citations; Kraken2: 524 citations) and a frequently used program, which means many researchers discuss problems, solutions, and approaches to using this software in online forums, such as Biostars.org. In addition to this, developers create third-party software for applications such as abundance estimation (using Bracken)(Lu et al. 2017) and classification visualisation (using Krona) (Ondov et al., 2011). Secondly, Kraken2 is efficient, only requiring a modest amount of time and computing power to run, additionally, it has a straightforward command line interface, making the creation of databases and classification less time-consuming. Furthermore, custom databases can be produced, allowing flexibility for many research applications, even allowing support for the ubiquitous bacterial 16S rRNA gene and protein databases. Other classification tools were considered for this application, including Centrifuge (Kim et al., 2016) and Kaiju (Menzel et al., 2015),

however, Kraken2 was considered superior based on the previously mentioned aspects (Breitwieser et al., 2019).

Though Kraken2 works well to isolate and classify reads to a taxonomic group, there is often a significant portion of reads that remain unclassified, which may be due to several factors. Firstly, the experimental NGS sample that has been produced may include sequences of DNA that have not been assigned to any genomes on the NCBI databases, as the Kraken2 classification depends greatly on readily available genomes from NCBI. This means that if sequences have not been uploaded as genomes yet, they will not find a match and will not be classified. Alternatively, the sequences that are not being classified may belong to genomes that are not included in the database used. This is why it is important to understand what the sample is comprised of and include any potential genomes in the Kraken2 database. Another possible explanation of high "No Hits" result like that of the dataset Lane 1 (where only 4.33% of the sequences reads classified as Ixodidae, when it was the majority of the extracted DNA sample) could be that the sequencing run failed for an un-detected reason, leading to sequence reads that are unusable.

The reliance of the Kraken2 classification also includes the most current recognition of species taxonomic classification, if all genomic content on NCBI is not yet up to date with the most current terminology, there may be error in the species names that reads are classified to. This can be seen when classifying the datasets Lane 1, *R. linnaei, R. appendiculatus* and *D. variabilis* where some reads have been classified into either *Rhipicephalus* or *Boophilus* even though they are now considered one species (Guglielmone et al., 2010). As long as these discrepancies are identified, they can be corrected during analysis and presentation of classification results.

Adding well-constructed genomes to Kraken2 databases does not have a significant effect on classification results. This was tested using the Kraken2 database 'tick+human-db' and was compared against the primary tick-focused 'tick-db'. The greatest variation in results was a -0.18% change in the highest classification result (Table 14) produced by the unassembled *R. appendiculatus*

dataset when comparing tick-db and tick+human-db. The average variation in classification percentage for the highest classification result was -0.025% across all datasets, including unassembled and assembled sequences. These findings suggest that too much caution is not essential when trying to produce a Kraken2 database with all possible genomes that could be in a shotgun sequencing sample; adding a potential genome (rather than excluding it for fear of skewing results) would be more beneficial. Despite this, it is recommended that genomes which are highly unlikely to be in the sample are not added, as this could result in false positives and an overly large database.

As with the other taxonomic isolation methods tested in the present study, understanding the characteristics of the organism used in sequencing is extremely beneficial as it allows for the identification of all possible genomes that could be present within the species. For example, ticks are known to harbour a wide range of microorganisms, including bacteria and virus. The species of microbes are likely to be detected using a default read classification database, like the Kraken2 'Standard' database, which includes bacteria, archaea, virus and human DNA. However, many tick species are also known to carry protozoan parasites, such as *Theileria* (Bishop et al., 2004; Marendy et al., 2020), *Babesia* (Blaschitz et al., 2008; Brinkmann et al., 2019) and potentially *Toxoplasma* (Ben-Harari, 2019; Kim et al., 2020). With protozoans being excluded in most default read classification databases, these pathogens will not be detected. It is therefore important to understand the target organism's characteristics and include the genomes that are not present in a default read classification database but could potentially be present.

When deciding whether assembled or unassembled reads are better for using in Kraken2 analysis, measuring the number of base pairs classified rather than percentage of reads classified is more informative for determining the level of genomic content recovered. This is because, the number of reads change depending on how the dataset is processed (i.e., using pure trimmed data will have more sequences than MEGAHIT assembled data), impacting the calculation of percentage classified.

The results from this study have shown that the number of base pairs classified changes when comparing assembled and unassembled data. The unassembled (quality controlled raw reads) having significantly more classified base pairs than MEGAHIT assembled data. Across all datasets, using MEGAHIT assembled contigs yielded an average of 88.2% less bases than using unfiltered quality-controlled sequences (Table 13). These results show that using an assembler will not aid in microsatellite primer development as less genomic content will be able to be recovered from a dataset. It is then not recommended to perform a genomic assembly before classifying reads using Kraken2 for the purpose of developing microsatellite markers.

Kraken2 was shown to be successful in classifying mixed-species Ixodida shotgun sequencing data in an improved method over using read alignment. Using Kraken2 for species classification is more computationally and time efficient compared to read alignment. As well as this, a Kraken2 classification only needs to be performed once, as reads are classified to multiple reference genomes in the Kraken2 database. In contrast, when performing read mapping, the alignment process must be performed multiple times against multiple reference genomes. However, it should be noted that Kraken2 and Bowtie2 computational requirements are different; Kraken2 is more memory taxing and therefore, a significant supply of RAM is required; conversely, Bowtie2 has a relatively low memory requirement but is more taxing on the CPU (Langmead & Salzberg, 2012; Wood et al., 2019).

Across all datasets, using Kraken2 for Ixodida classification produced 617 primer sets created using Krait and Primer3. This was a small increase compared to using unfiltered trimmed data, which produced 578 primers sets. This is an unexpected decrease as there were less sequences to be scanned for microsatellites as only the Ixodida classified reads were kept. This could be the result of either a bug in the code of Krait, which has not been corrected or an inconsistency in the data that has not been detected. However, the change between untrimmed and Kraken2 classified data is

small and as the data has been classified as Ixodidae by Kraken2, the microsatellite primers developed from these datasets can be presumed to come from tick.

4.5 - Microsatellite Detection and Primer Development

Due to the nature of this project, primer development was fast-tracked to have a significant portion of time to test them against *I. holocyclus* DNA in the wet lab. Consequently, the primers were developed from trimmed un-isolated reads before the discovery that the dataset was not appropriate for downstream analysis. Therefore, after some unsuccessful attempts, experimentation of the developed primers was abandoned in favour of focusing on finding the best methods of read isolation and the formation of a workflow. Primers were initially screened using a standard KAPA Taq PCR protocol to create a baseline that could be adjusted as necessary. As it was becoming clear that there was a low likelihood that the developed primers would be successful, a second PCR experiment was performed, which included an increase of magnesium added to the reactions to increase sensitivity. However, this did not produce amplification and the experimentation of these primers was abandoned.

One of the most favourable STR detection programs discovered in the project was Krait. The main advantages of Krait are its ease of use (because of the graphic user interface), as well as it being available for Windows, macOS and Linux (Du et al., 2018). Krait also contains built-in microsatellite filtering tools that allow the user to search for microsatellites according to their desired parameters once all potential microsatellites have been discovered. In addition to this, Krait has the Primer3 tool built into the program to allow time-efficient primer development. A graphical interface for bioinformatics programs is reasonably rare, so this is a significant advantage for people who are not familiar with systems like UNIX. The graphical interface also has drawbacks in that it is less efficient for parsing data, requiring modest CPU power and the only powerful computing systems many researchers have access to are command-line Linux systems, meaning that the program is not able

to be used. However, in discussions with the developer of Krait on GitHub, it has been reported that a command-line version of the Krait program is currently under development (Du, 2020/2021).

In 2020, Cai et al. used the program Krait to scan *Rhinopithecus roxellana* (snub-nosed monkey) genomic content for microsatellites for the purpose of creating a test that can perform individual identification and paternity testing. However, in contrast to this study, the snub-nosed monkey is a highly threatened species and therefore has multiple assembled genomes for research use (NCBI: 4 reference genomes). The benefit to having a well-constructed genome is there are few limitations to the size of the microsatellite flanking region that is to be recovered. This is in contrast to short-read NGS data where many microsatellite sequences are too long to be able to generate primer sets for either side of the microsatellite region. This allowed the researchers to parse a well assembled genome to scan for microsatellites. Yet, many organisms of research interest do not have the access to a well assembled reference genome making the methods used in this study a practical option for less documented species (Cai et al., 2020).

Previous studies that have developed microsatellite primers for ticks used much more laborious, costly methods to discover microsatellites that have the potential to be population markers. Many papers follow a protocol that involves variations of the following steps: target population DNA extraction, digestion of DNA with restriction enzymes, ligation of DNA linkers, DNA fragments are separated by electrophoresis, fragments of a certain size are extracted from the gel, filtered fragments are enriched for specific microsatellite motifs, DNA is hybridised to repeat oligos, DNA samples are then washed, enriched fragments are recovered, fragments are transformed into *E.coli*, *E. coli* are screened using probes, positive clones are sequenced, primers are designed for appropriate microsatellite sequences, primers are tested on target DNA (Gardner et al., 2008; Guzinski et al., 2008; Koffi et al., 2006). The success rate for the cost of the wet lab workflow is also low, with 14 microsatellite markers out of 259 *E.coli* positives for *Dermacentor albipictus* (Leo et al.,

2012) and 10 out of the 20 positives in finding markers for *Bothriocroton hydrosauri* (Guzinski et al., 2008) being created and used.

Other papers used hybrid approaches, where DNA was fragmented and enriched for microsatellites motifs, then sequenced using NGS, whereby reads were assembled and searched for microsatellites using an STR detection program (Van Houtte et al., 2013; Van Oosten et al., 2014). Using this method still required the fragmentation and enrichment of DNA for microsatellite motifs, while the workflow presented in this thesis works on shotgun mixed species sample, allowing the data to be used for other applications besides the development of microsatellite markers. Compared to this labour intensive and resource costly process, using the bioinformatics workflow described in this thesis, all that is required is shotgun sequencing of the target population, sequence data to be piped through the presented in silico workflow and primer testing on individuals of the target population.

Of the different read isolation methods that were experimented (MEGAHIT assembled contigs, Bowtie2 filtered reads and Kraken2 filtered reads), the method that produced the most developed microsatellite primers was the MEGAHIT assembly, producing a total of 2792 primer sets compared to the pure quality controlled reads which produced 578 primer sets. This increase in primer sets generated is likely due to the benefit of having longer contigs, allowing larger microsatellite regions to be detected my Krait and larger flanking regions so Primer3 can more easily generate primer sets. However, many reads are discarded due to not being assembled and these reads may contain microsatellites. Kraken2 produced slightly more microsatellite primers sets with a total of 617 generated. This result is unexpected as fewer sequence reads (due to non-tick content being removed from analysis) were scanned for microsatellites by Krait. The primer sets generated using Bowtie2 for read isolation had an expected decrease (441), as there were less sequence reads for Krait to scan due to sequences being removed that did not map to a tick reference genome.

From these results it is recommend, for the purposes of obtaining microsatellite markers from short read NGS data, that sequence reads are classified by a custom Kraken2 database to isolate target

genomic content, then isolated sequences are assembled. Alternatively, if large volumes of RAM are not available, performing a read alignment on a closely related genome will isolate a sufficient number of sequences. Once sequences are isolated, they can be scanned for microsatellites and primers can be created for the flanking regions using the researcher's program of choice.

4.6 - Bioinformatics Software and Workflows

Open-source software are programs, scripts, code etc., are made freely available to the general public, allowing free use of the software and free use to modify the software as desired. This allows the public to help develop and improve the software to the needs of the community, creating ideal programs (Stajich & Lapp, 2006). Due to this, open-source programs are excellent for scientific applications, where transparency and the ability to experiment with different variables are vitally important. Many bioinformatics programs available are open-source, allowing researchers to use the programs without cost, experiment with programs that best fit their project, and, if required, make changes to the source code. The programs used in the pipeline developed in this thesis are, for the most part open source, allowing researchers to use and modify.

GitHub.com is an important platform for many programmers, allowing code to be published, distributed and edited by collaborators in a streamlined way (Seker et al., 2020). The majority of bioinformatics programs currently being developed are uploaded to GitHub to allow public access, and to test and provide critical evaluation on the programs being developed (Perez-Riverol et al., 2016). The use of GitHub allows almost instant use of newly developed programs that have improvements over previously used programs. In addition to this, trends in community usage of a program can allow the discovery of desirable programs for a project (Dozmorov, 2018). The BASH scripts that have been produced for this thesis are available on GitHub (https://github.com/XWBarton/HonoursScripts), so they can be used, analysed and changed by any user. These scripts that have been created use other open-source developed software, including Kraken2, Bowtie2, Bracken, Krona and Samtools, and other inbuilt UNIX programs.

4.7 - Limitations

During this project, one concerning occurrence is the differences observed in total read mapping percentage between the different read mapping programs. These experiments (*see* Table 11 and Figures 2 and 3) show that bowtie2 is significantly more stringent in read mapping than BWA-MEM despite using the very sensitive local setting. This creates a concern for using BWA-MEM for this purpose as a large generation of false-positive sequences may contain microsatellite sequences that primers will be developed for, which is not time or cost efficient for testing in a wet lab, because of this, bowtie2 was favoured for this workflow.

This also creates a more general concern for lack of controls and tests to create a benchmark on available bioinformatics programs, making it difficult to decide which program is best for a project. For more common programs (e.g. read mappers and genome assemblers), there are at least peer-reviewed comparisons measuring the results on artificial input data (Thankaswamy-Kosalai et al., 2017; Forouzan et al., 2018; Khan et al., 2018), yet for less-known programs with more niche applications, such as STR detection or taxon classification, there are very few current benchmarks that have been completed to view the performance of the software and provide a comparison of results between different programs. This leads to researchers having to decide what program to use for a project by reading consensus amongst other researchers in non-peer reviewed work, such as websites, blogs, and forums.

When trying to discover optimal programs for a project, it was often the case that programs are no longer maintained or were simply unavailable for use due to the servers that the code was stored on being shut down. For future programs being developed, it is suggested that uploading the source code to platforms, such as GitHub, without the need for constant maintenance, which allows researchers to use the code in the long term. As well as this, many programs developed contain minimal documentation on how a program is used and what its benefits and limitations are. If there is no documentation when programs are no longer supported, there is often no way of successfully

using the program, making the time and effort someone has put in the development of the program redundant.

Another limitation of this study was the initial data received from AGRF. The Illumina NovaSeq 6000 Data had initial promising results according to the fastp report, with the general quality of the reads being high, as all sequences had a quality score above 33 and base contents being evenly distributed before filtering. Once fastp trimming occurred with all possible poor reads removed, the data continued to appear of good quality. Nevertheless, once the data was aligned to other genomes, only minimal reads were mapped and classified. When mapping with bowtie2, around 2% of the reads could be aligned across all Ixodidae reference genomes. When mapping with BWA-MEM, significantly more reads were mapped, yet there was very little discrimination between the reference genomes with the exception of *E. coli*. These mapping results were in contrast to other datasets, which had a discriminately best mapped reference genome. As well as this, the read classification results using Kraken2 and tick-db were extremely low at only 4.43% of the reads being classified as Ixodidae, contrasting with the other datasets that had a minimum Ixodidae classification of 31.3% (*D. variabilis*).

Unique to this project was the use of non-microbial metagenomic sequencing samples that contain not only microbial content such as bacteria, protozoans and virus, but also Ixodida arthropod content. Many of the pipelines and tools which are currently available are tailored explicitly to microbe only datasets. Due to this, the comparison of these tools and pipelines in peer-reviewed literature only use microbial datasets to assess the quality of the workflows, leaving a gap in knowledge for tools that work with non-microbial content (Yue et al., 2020).

To develop microsatellite markers for future studies, it is strongly recommended that long sequencing techniques, such as PacBio must be used. Short sequence reads like the sequence data used in this project are significantly too short to be able to contain many of the microsatellite repeats found in a genome, along with the flanking regions for which primers must be developed.

Being able to use long-read sequencing will allow many entire microsatellite regions to be contained within single reads along with their associated flanking regions. This contrasts to the many sequences which are lost in short-read sequencing, as using assembly and read mapping methods to piece back together only works effectively for unique sequences. As microsatellites are in their nature generic, it is a great challenge to produce correct microsatellite regions out of multiple sequences. As long-read sequencing technologies are recently becoming less costly, and more importantly, increasing in accuracy, implementation of this technology is highly recommended.

Finally, these findings were limited by not being able to confirm the success of the microsatellite primers that were developed from the bioinformatics workflow due to the available data not performing as expected. Until PCR testing of these developed primers can be performed, it is uncertain whether this workflow works in its entirety.

4.8 - Future Directions

In future implementations of this workflow, instead of performing multiple alignment runs of Bowtie2 which is not time efficient, it is recommended to use a program such as FastQ Screen (Wingett & Andrews, 2018). FastQ Screen is a time efficient program that allows read mapping of a dataset to multiple reference genomes, in order to confirm the origin of an organism by using a read mapping program such as Bowtie2 or BWA. It does this by automatically aligning the query datasets to a chosen range of reference genomes, producing a summary of mapping results, allowing a user to find the best-mapped genome, isolate reads, assess if a sequencing run has been successful and potentially identify the organism of the sequencing run.

Due to some of the difficulties in developing population genetics microsatellite markers, SNP markers are becoming increasingly popular due to their ease of discovery and analyses (Araya-Anchetta et al., 2015). SNPs are an excellent way to measure genetic variation in non-model organisms such as ticks, being a direct measure of genetic diversity. An effective way to analyse genetic variation in a group of organisms such as ticks, would be to develop a custom “SNP-chip”, a

type of DNA micro-array containing a selection of SNPs from a target organism genetic sequence. This SNP-chip design can then be used to measure the genetic diversity of organism in a population (Hagen et al., 2013).

Lastly, in future studies relating to the development of population genetics microsatellite markers, it is imperative that the DNA that is used to test the created microsatellite markers comes from the same population that the sample organism came from that underwent shotgun sequencing. This way, new microsatellite markers have a higher chance of successful amplification.

4.9 - Conclusion

The main application of this project was to be able to create population genetic markers using microsatellite repeat region for *I. holocyclus*, generated from an *I. holocyclus* whole-body shotgun dataset, processed by a developed bioinformatics workflow. With the intention of creating successful markers for the analysis of genetic differences in tick populations in a geographical area and through time, as well as a measure of inbreeding populations, subpopulations and diseases associated with specific species of ticks.

Determining the quality of NGS run is vital to the success of a project and should first be completed with programs such as FastQC, Trimmomatic or fastp. For the majority of sequencing runs, this will reveal whether sequencing has been successful and can be used for downstream analysis. However, a dataset may still have failed in some way that is not detected by these quality checking programs. This means that, in addition to this, identifying how well a dataset is mapped to a closely related genome, how it is classified according to a tailored taxon classification database and what contigs are generated by an assembler can indicate whether a sequencing run has succeeded and whether there is high-quality data for analysis.

Being able to isolate different taxon from a mixed-species shotgun next generation run appears to be feasible, and the recommended workflow for this would be using a combination of read alignment software, read taxon classification software and sequence assembly. Using these

bioinformatics tools has shown to be useful in separating tick genomic content from other genomes that may be in the sequencing run. Despite the focus of many metagenomics studies being on microbial communities, the metagenomic bioinformatics programs used in this project appeared to perform as intended for isolating Ixodida DNA sequences.

From the results obtained from this study, producing microsatellite markers from short-read NGS datasets is possible, though it is not optimal for two reasons. Firstly, short-read NGS are often too short to contain the entire length of a microsatellite region, meaning that many flanking regions are not covered, and primers for these regions cannot be created. Secondly, microsatellite regions are, by their nature, generic and therefore using different techniques to organise short-read NGS data (i.e., genome assembly, read alignment or read taxon classification) leaves out many microsatellite regions as they are not able to be uniquely placed. For these reasons, long-read sequencing would be strongly recommended for future projects.

The recommended workflow and associated software described in this thesis will allow future researchers to more easily decide on what approach to take in generating microsatellites from mixed species or even pure shotgun NGS data, reducing costs in time, money and resources.

5 - References

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Andreotti, R., Cunha, R. C., Soares, M. A., Guerrero, F. D., Leivas Leite, F. P., & Pérez de León, A. A. (2012). Protective immunity against tick infestation in cattle vaccinated with recombinant trypsin inhibitor of *Rhipicephalus microplus*. *Vaccine*, *30*(47), 6678–6685. https://doi.org/10.1016/j.vaccine.2012.08.066

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Araya-Anchetta, A., Busch, J. D., Scoles, G. A., & Wagner, D. M. (2015). Thirty years of tick population genetics: A comprehensive review. *Infection, Genetics and Evolution*, *29*, 164–179. https://doi.org/10.1016/j.meegid.2014.11.008

Araya-Anchetta, A., Scoles, G. A., Giles, J., Busch, J. D., & Wagner, D. M. (2013). Hybridization in natural sympatric populations of *Dermacentor* ticks in northwestern North America. *Ecology and Evolution*, *3*(3), 714–724. https://doi.org/10.1002/ece3.496

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barbosa, A., Benzal, J., Vidal, V., D'Amico, V., Coria, N., Diaz, J., Motas, M., Palacios, M. J., Cuervo, J. J., Ortiz, J., & Chitimia, L. (2011). Seabird ticks (*Ixodes uriae*) distribution along the Antarctic Peninsula. *Polar Biology*, *34*(10), 1621–1624. https://doi.org/10.1007/s00300-011-1000-7

Barker, D. (2019). Ixodes barkeri n. sp. (Acari: Ixodidae) from the short-beaked echidna, *Tachyglossus aculeatus* , with a revised key to the male *Ixodes* of Australia, and list of the subgenera and species of *Ixodes* known to occur in Australia. *Zootaxa*, *4658*(2), 331–342. https://doi.org/10.11646/zootaxa.4658.2.7

Barker, S. C., & Walker, A. R. (2014). Ticks of Australia: The species that infest domestic animals and humans. Magnolia Press.

Barker, S. C., Walker, A. R., & Campelo, D. (2014). A list of the 70 species of Australian ticks; diagnostic guides to and species accounts of *Ixodes holocyclus* (paralysis tick), *Ixodes cornuatus* (southern paralysis tick) and *Rhipicephalus australis* (Australian cattle tick); and consideration of the place of Australia in the evolution of ticks with comments on four controversial ideas. *International Journal for Parasitology*, *44*(12), 941–953. https://doi.org/10.1016/j.ijpara.2014.08.008

Barrero, R. A., Guerrero, F. D., Black, M., McCooke, J., Chapman, B., Schilkey, F., Pérez de León, A. A., Miller, R. J., Bruns, S., Dobry, J., Mikhaylenko, G., Stormo, K., Bell, C., Tao, Q., Bogden, R., Moolhuijzen, P. M., Hunter, A., & Bellgard, M. I. (2017). Gene-enriched draft genome of the cattle tick *Rhipicephalus microplus*: Assembly by the hybrid Pacific Biosciences/Illumina approach enabled analysis of the highly repetitive genome. *International Journal for Parasitology*, *47*(9), 569–583. https://doi.org/10.1016/j.ijpara.2017.03.007

Beati, L., & Keirans, J. E. (2001). Analysis of the Systematic Relationships among Ticks of the Genera *Rhipicephalus* and *Boophilus* (Acari: Ixodidae) Based on Mitochondrial 12S Ribosomal DNA Gene Sequences and Morphological Characters. *The Journal of Parasitology*, *87*(1), 32–48. https://doi.org/10.2307/3285173

Bendele, K. G., Guerrero, F. D., Cameron, C., Bodine, D. M., & Miller, R. J. (2019). Gene expression during the early stages of host perception and attachment in adult female *Rhipicephalus microplus* ticks. *Experimental and Applied Acarology*, *79*(1), 107–124. https://doi.org/10.1007/s10493-019-00420-1

Ben-Harari, R. R. (2019). Tick transmission of toxoplasmosis. *Expert Review of Anti-Infective Therapy*, *17*(11), 911–917. https://doi.org/10.1080/14787210.2019.1682550

Bishop, R., Musoke, A., Morzaria, S., Gardner, M., & Nene, V. (2004). *Theileria*: Intracellular protozoan parasites of wild and domestic ruminants transmitted by ixodid ticks. *Parasitology*, *129 Suppl*, S271-83. https://doi.org/10.1017/S0031182003004748

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., & Taylor, J. (2010). Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*, *89*(1), 19.10.1-19.10.21. https://doi.org/10.1002/0471142727.mb1910s89

Blaschitz, M., Narodoslavsky-Gföller, M., Kanzler, M., Stanek, G., & Walochnik, J. (2008). *Babesia* Species Occurring in Austrian Ixodes ricinus Ticks. *Applied and Environmental Microbiology*, *74*(15), 4841–4846. https://doi.org/10.1128/AEM.00035-08

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., … Korf, I. F. (2013). Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(2047-217X-2–10). https://doi.org/10.1186/2047-217X-2-10

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1136. https://doi.org/10.1093/bib/bbx120

Brinkmann, A., Hekimoğlu, O., Dinçer, E., Hagedorn, P., Nitsche, A., & Ergünay, K. (2019). A cross-sectional screening by next-generation sequencing reveals *Rickettsia*, *Coxiella*, *Francisella*, *Borrelia*, *Babesia*, *Theileria* and *Hemolivia* species in ticks from Anatolia. *Parasites & Vectors*, *12*(1), 26. https://doi.org/10.1186/s13071-018-3277-7

Buehler, A. J., Evanowski, R. L., Martin, N. H., Boor, K. J., & Wiedmann, M. (2017). Internal transcribed spacer (ITS) sequencing reveals considerable fungal diversity in dairy products. *Journal of Dairy Science*, *100*(11), 8814–8825. https://doi.org/10.3168/jds.2017-12635

Buettner, P. G., Westcott, D. A., Maclean, J., Brown, L., McKeown, A., Johnson, A., Wilson, K., Blair, D., Luly, J., Skerratt, L., Muller, R., & Speare, R. (2013). Tick Paralysis in Spectacled Flying-Foxes (*Pteropus conspicillatus*) in North Queensland, Australia: Impact of a Ground-Dwelling Ectoparasite Finding an Arboreal Host. *PLOS ONE*, *8*(9), e73078. https://doi.org/10.1371/journal.pone.0073078

Busch, J. D., Stone, N. E., Nottingham, R., Araya-Anchetta, A., Lewis, J., Hochhalter, C., Giles, J. R., Gruendike, J., Freeman, J., Buckmeier, G., Bodine, D., Duhaime, R., Miller, R. J., Davey, R. B., Olafson, P. U., Scoles, G. A., & Wagner, D. M. (2014). Widespread movement of invasive cattle fever ticks (*Rhipicephalus microplus*) in southern Texas leads to shared local infestations on cattle and deer. *Parasites & Vectors*, *7*(1), 188. https://doi.org/10.1186/1756-3305-7-188

Butler, J. M. (2012). Chapter 16—Non-human DNA. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Methodology* (pp. 473–495). Academic Press. https://doi.org/10.1016/B978-0-12-374513-2.00016-6

Cai, Y., Yu, H., Liu, H., Jiang, C., Sun, L., Niu, L., Liu, X., Li, D., & Li, J. (2020). Genome-wide screening of microsatellites in golden snub-nosed monkey (*Rhinopithecus roxellana*), for the development of a standardized genetic marker system. *Scientific Reports*, *10*(1), 10614. https://doi.org/10.1038/s41598-020-67451-2

Cao, Y., Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01829

Carpi, G., Cagnacci, F., Wittekindt, N. E., Zhao, F., Qi, J., Tomsho, L. P., Drautz, D. I., Rizzoli, A., & Schuster, S. C. (2011). Metagenomic Profile of the Bacterial Communities Associated with

Ixodes ricinus Ticks. *PLOS ONE*, *6*(10), e25604. https://doi.org/10.1371/journal.pone.0025604

Casillas, S., & Barbadilla, A. (2017). Molecular Population Genetics. *Genetics*, *205*(3), 1003–1035. https://doi.org/10.1534/genetics.116.196493

Chand, K. K., Lee, K. M., Lavidis, N. A., Rodriguez-Valle, M., Ijaz, H., Koehbach, J., Clark, R. J., Lew-Tabor, A., & Noakes, P. G. (2016). Tick holocyclotoxins trigger host paralysis by presynaptic inhibition. *Scientific Reports*, *6*. https://doi.org/10.1038/srep29446

Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., Wu, J., Yang, H., Gao, J., & Huang, L. (2021). Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nature Communications*, *12*(1), 1106. https://doi.org/10.1038/s41467-021-21295-0

Chen, K., & Pachter, L. (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLOS Computational Biology*, *1*(2), e24. https://doi.org/10.1371/journal.pcbi.0010024

Chen, M., & Zhao, H. (2019). Next-generation sequencing in liquid biopsy: Cancer screening and early detection. *Human Genomics*, *13*(1), 34. https://doi.org/10.1186/s40246-019-0220-8

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chudleigh, F., & Franco-Dixon, M. A. (2010). An economic evaluation of tick line deregulation in Queensland (No. 421-2016–26778). *Australian Agricultural and Resource Economics Society*. https://doi.org/10.22004/ag.econ.58889

Commins, S. P., & Platts-Mills, T. A. E. (2013). Delayed Anaphylaxis to Red Meat in Patients with IgE Specific for Galactose alpha-1,3-Galactose (alpha-gal). *Current Allergy and Asthma Reports*, *13*(1), 72–77. https://doi.org/10.1007/s11882-012-0315-y

Cooper, A., Stephens, J., Ketheesan, N., & Govan, B. (2012). Detection of *Coxiella burnetii* DNA in Wildlife and Ticks in Northern Queensland, Australia. *Vector-Borne and Zoonotic Diseases*, *13*(1), 12–16. https://doi.org/10.1089/vbz.2011.0853

Cooper, B. J., & Spence, I. (1976). Temperature-dependent inhibition of evoked acetylcholine release in tick paralysis. *Nature*, *263*(5579), 693–695. https://doi.org/10.1038/263693a0

Cutullé, C., Jonsson, N. N., & Seddon, J. (2009). Population structure of Australian isolates of the cattle tick *Rhipicephalus* (*Boophilus*) *microplus*. *Veterinary Parasitology*, *161*(3), 283–291. https://doi.org/10.1016/j.vetpar.2009.01.005

Cutullé, C., Jonsson, N. N., & Seddon, J. M. (2010). Multiple paternity in *Rhipicephalus* (*Boophilus*) *microplus* confirmed by microsatellite analysis. *Experimental & Applied Acarology; Dordrecht*, *50*(1), 51–58. http://dx.doi.org.libproxy.murdoch.edu.au/10.1007/s10493-009-9298-3

de la Fuente, J. (2003). The fossil record and the origin of ticks (Acari: Parasitiformes: Ixodida). *Experimental & Applied Acarology*, *29*(3), 331–344. https://doi.org/10.1023/A:1025824702816

Dehhaghi, M., Kazemi Shariat Panahi, H., Holmes, E. C., Hudson, B. J., Schloeffel, R., & Guillemin, G. J. (2019). Human Tick-Borne Diseases in Australia. *Frontiers in Cellular and Infection Microbiology*, *9*. https://doi.org/10.3389/fcimb.2019.00003

Derrick, E. H., Smith, D. J. W., & Brown, H. E. (1942). Studies in the Epidemiology of Q Fever. *Australian Journal of Experimental Biology and Medical Science*, *20*(2), 105–110. https://doi.org/10.1038/icb.1942.19

Díaz-Sánchez, S., Hernández-Jarguín, A., Torina, A., de Mera, I. G. F., Blanda, V., Caracappa, S., Gortazar, C., & de la Fuente, J. (2019). Characterization of the bacterial microbiota in wild-caught *Ixodes ventalloi*. *Ticks and Tick-Borne Diseases*, *10*(2), 336–343. https://doi.org/10.1016/j.ttbdis.2018.11.014

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dozmorov, M. G. (2018). GitHub Statistics as a Measure of the Impact of Open-Source Bioinformatics Software. *Frontiers in Bioengineering and Biotechnology*, *6*. https://doi.org/10.3389/fbioe.2018.00198

Du, L. (2021). *Lmdu/stria* [C]. https://github.com/lmdu/stria (Original work published 2020)

Du, L., Zhang, C., Liu, Q., Zhang, X., & Yue, B. (2018). Krait: An ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics*, *34*(4), 681–683. https://doi.org/10.1093/bioinformatics/btx665

Duron, O., Sidi-Boumedine, K., Rousset, E., Moutailler, S., & Jourdain, E. (2015). The Importance of Ticks in Q Fever Transmission: What Has (and Has Not) Been Demonstrated? *Trends in Parasitology*, *31*(11), 536–552. https://doi.org/10.1016/j.pt.2015.06.014

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. Scopus. https://doi.org/10.1093/bioinformatics/btq461

Egan, S. (2017). Profiling the bacterial microbiome of ticks that parasitise bandicoots in Australia [Honours, Murdoch University]. In *Egan, Siobhon <https://researchrepository.murdoch.edu.au/view/author/Egan, Siobhon.html>ORCID: 0000-0003-4395-4069 <http://orcid.org/0000-0003-4395-4069> (2017) Profiling the bacterial microbiome of ticks that parasitise bandicoots in Australia. Honours thesis, Murdoch University.* https://researchrepository.murdoch.edu.au/id/eprint/40003/

Egan, S. L., Loh, S.-M., Banks, P. B., Gillett, A., Ahlstrom, L., Ryan, U. M., Irwin, P. J., & Oskam, C. L. (2020). Bacterial community profiling highlights complex diversity and novel organisms in wildlife ticks. *Ticks and Tick-Borne Diseases*, *11*(3), 101407. https://doi.org/10.1016/j.ttbdis.2020.101407

Eppleston, K. R., Kelman, M., & Ward, M. P. (2013). Distribution, seasonality and risk factors for tick paralysis in Australian dogs and cats. *Veterinary Parasitology*, *196*(3), 460–468. https://doi.org/10.1016/j.vetpar.2013.04.011

Evans, M. (2018). Molecular barcoding of Australian ticks [Honours, Murdoch University]. In *Evans, Megan <https://researchrepository.murdoch.edu.au/view/author/Evans, Megan.html> (2018) Molecular barcoding of Australian ticks. Honours thesis, Murdoch University.* https://researchrepository.murdoch.edu.au/id/eprint/42887/

Evans, M. L., Egan, S., Irwin, P. J., & Oskam, C. L. (2019). Automatic Barcode Gap Discovery reveals large COI intraspecific divergence in Australian Ixodidae. *Zootaxa*, *4656*(2), 393–396. https://doi.org/10.11646/zootaxa.4656.2.13

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fagerberg, A. J., Fulton, R. E., & Black, W. C. (2001). Microsatellite loci are not abundant in all arthropod genomes: Analyses in the hard tick, *Ixodes scapularis* and the yellow fever mosquito, *Aedes aegypti*. *Insect Molecular Biology*, *10*(3), 225–236. https://doi.org/10.1046/j.1365-2583.2001.00260.x

Faircloth, B. C. (2008). msatcommander: Detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, *8*(1), 92–94. https://doi.org/10.1111/j.1471-8286.2007.01884.x

Fan, J., Huang, S., & Chorlton, S. D. (2021). BugSeq: A highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics*, *22*(1), 160. https://doi.org/10.1186/s12859-021-04089-5

Forouzan, E., Shariati, P., Mousavi Maleki, M. S., Karkhane, A. A., & Yakhchali, B. (2018). Practical evaluation of 11 de novo assemblers in metagenome assembly. *Journal of Microbiological Methods*, *151*, 99–105. https://doi.org/10.1016/j.mimet.2018.06.007

Foulkes, A. C., Watson, D. S., Griffiths, C. E. M., Warren, R. B., Huber, W., & Barnes, M. R. (2017). Research Techniques Made Simple: Bioinformatics for Genome-Scale Biology. *Journal of Investigative Dermatology*, *137*(9), e163–e168. https://doi.org/10.1016/j.jid.2017.07.095

Fungtammasan, A., Ananda, G., Hile, S. E., Su, M. S.-W., Sun, C., Harris, R., Medvedev, P., Eckert, K., & Makova, K. D. (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research*, *25*(5), 736–749. https://doi.org/10.1101/gr.185892.114

Gardner, M. G., Sanchez, J. J., Dudaniec, R. Y., Rheinberger, L., Smith, A. L., & Saint, K. M. (2008). *Tiliqua rugosa* microsatellites: Isolation via enrichment and characterisation of loci for multiplex PCR in *T. rugosa* and the endangered *T. adelaidensis*. *Conservation Genetics*, *9*(1), 233–237. https://doi.org/10.1007/s10592-007-9316-0

Gemmell, R. T., Cepon, G., Green, P. E., & Stewart, N. P. (1991). SOME EFFECTS OF TICK INFESTATIONS ON JUVENILE NORTHERN BROWN BANDICOOT (*ISOODON MACROURUS*). *Journal of Wildlife Diseases*, *27*(2), 269–275. https://doi.org/10.7589/0090-3558-27.2.269

Grattan-Smith, P. J., Morris, J. G., Johnston, H. M., Yiannikas, C., Malik, R., Russell, R., & Ouvrier, R. A. (1997). Clinical and neurophysiological features of tick paralysis. *Brain*, *120*(11), 1975–1987. https://doi.org/10.1093/brain/120.11.1975

Graves, S. R., Jackson, C., Hussain-Yusuf, H., Vincent, G., Nguyen, C., Stenos, J., & Webster, M. (2016). *Ixodes holocyclus* Tick-Transmitted Human Pathogens in North-Eastern New South Wales, Australia. *Tropical Medicine and Infectious Disease*, *1*(1), 4. https://doi.org/10.3390/tropicalmed1010004

Graves, S. R., & Stenos, J. (2017). Tick-borne infectious diseases in Australia. *Medical Journal of Australia*, *206*(7), 320–324. https://doi.org/10.5694/mja17.00090

Greay, T. L., Gofton, A. W., Paparini, A., Ryan, U. M., Oskam, C. L., & Irwin, P. J. (2018). Recent insights into the tick microbiome gained through next-generation sequencing. *Parasites & Vectors*, *11*(1), 12. https://doi.org/10.1186/s13071-017-2550-5

Greay, T. L., Oskam, C. L., Gofton, A. W., Rees, R. L., Ryan, U. M., & Irwin, P. J. (2016). A survey of ticks (Acari: Ixodidae) of companion animals in Australia. *Parasites & Vectors*, *9*(1), 207. https://doi.org/10.1186/s13071-016-1480-y

Gregory, T. R., & Young, M. R. (2020). Small genomes in most mites (but not ticks). *International Journal of Acarology*, *46*(1), 1–8. https://doi.org/10.1080/01647954.2019.1684561

Guerrero, F. D., Andreotti, R., Bendele, K. G., Cunha, R. C., Miller, R. J., Yeater, K., & Pérez de León, A. A. (2014). *Rhipicephalus* (*Boophilus*) *microplus* aquaporin as an effective vaccine antigen to protect against cattle tick infestations. *Parasites & Vectors*, *7*, 475. https://doi.org/10.1186/s13071-014-0475-9

Guglielmone, A. A., Petney, T. N., & Robbins, R. G. (2020). Ixodidae (Acari: Ixodoidea): descriptions and redescriptions of all known species from 1758 to December 31, 2019. *Zootaxa*, *4871*(1), zootaxa.4871.1.1. https://doi.org/10.11646/zootaxa.4871.1.1

Guglielmone, A. A., Robbins, R. G., Apanaskevich, D. A., Petney, T. N., Estrada-Peña, A., Horak, I. G., Shao, R., & Barker, S. C. (2010). The Argasidae, Ixodidae and Nuttalliellidae (Acari: Ixodida) of the world: a list of valid species names. *Zootaxa*, *2528*(1), 1. https://doi.org/10.11646/zootaxa.2528.1.1

Guzinski, J., Saint, K. M., Gardner, M. G., Donnellan, S. C., & Bull, C. M. (2008). PERMANENT GENETIC RESOURCES: Development of microsatellite markers and analysis of their inheritance in the Australian reptile tick, *Bothriocroton hydrosauri*. *Molecular Ecology Resources*, *8*(2), 443–445. https://doi.org/10.1111/j.1471-8286.2007.01987.x

Hagen, I. J., Billing, A. M., Rønning, B., Pedersen, S. A., Pärn, H., Slate, J., & Jensen, H. (2013). The easy road to genome-wide medium density SNP screening in a non-model species: Development and application of a 10 K SNP-chip for the house sparrow (*Passer domesticus*). *Molecular Ecology Resources*, *13*(3), 429–439. https://doi.org/10.1111/1755-0998.12088

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R., & Nielsen, E. E. (2011). Application of SNPs for

population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, *11*(s1), 123–136. https://doi.org/10.1111/j.1755-0998.2010.02943.x

Hernandez, R., Chen, A. C., Davey, R. B., Ivie, G. W., Wagner, G. G., & George, J. E. (1998). Comparison of Genomic DNA in Various Strains of *Boophilus microplus* (Acari: Ixodidae). *Journal of Medical Entomology*, *35*(5), 895–900. https://doi.org/10.1093/jmedent/35.5.895

Horak, I. G., Camicas, J.-L., & Keirans, J. E. (2003). The Argasidae, Ixodidae and Nuttalliellidae (Acari: Ixodida): A World List of Valid Tick Names. 28, 27–54. *Experimental & Applied Acarology. https://doi-org.libproxy.murdoch.edu.au/10.1023/A:1025381712339*

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*. https://doi.org/10.1016/j.humimm.2021.02.012

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., & Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, *24*(4), 688–696. https://doi.org/10.1101/gr.168450.113

Irwin, P. J., Robertson, I. D., Westman, M. E., Perkins, M., & Straubinger, R. K. (2017). Searching for Lyme borreliosis in Australia: Results of a canine sentinel study. *Parasites & Vectors*, *10*(1), 114. https://doi.org/10.1186/s13071-017-2058-z

Jackson, J., Beveridge, I., Chilton, N., & Andrews, R. (2007). Distributions of the paralysis ticks *Ixodes cornuatus* and *Ixodes holocyclus* in south-eastern Australia. *Australian Veterinary Journal*, *85*(10), 420–424. https://doi.org/10.1111/j.1751-0813.2007.00183.x

Jang, J., Hur, H.-G., Sadowsky, M. J., Byappanahalli, M. N., Yan, T., & Ishii, S. (2017). Environmental Escherichia coli: Ecology and public health implications—a review. *Journal of Applied Microbiology*, *123*(3), 570–581. https://doi.org/10.1111/jam.13468

Jongejan, F., & Uilenberg, G. (2004). The global importance of ticks. *Parasitology*, *129*(S1), S3–S14. https://doi.org/10.1017/S0031182004005967

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N., & Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evolutionary Bioinformatics*, *14*, 1176934318758650. https://doi.org/10.1177/1176934318758650

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. Scopus. https://doi.org/10.1038/s41587-019-0201-4

Kim, Daehwan, Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Kim, J. Y., Kwak, Y. S., Lee, I.-Y., & Yong, T.-S. (2020). Molecular Detection of *Toxoplasma Gondii* in *Haemaphysalis* Ticks in Korea. *The Korean Journal of Parasitology*, *58*(3), 327–331. https://doi.org/10.3347/kjp.2020.58.3.327

Koffi, B. B., Risterucci, A. M., Joulia, D., Durand, P., Barré, N., Meeûs, T. D., & Chevillon, C. (2006). Characterization of polymorphic microsatellite loci within a young *Boophilus microplus* metapopulation. *Molecular Ecology Notes*, *6*(2), 502–504. https://doi.org/10.1111/j.1471-8286.2006.01295.x

Kwak, M. L., & Madden, C. (2017). The first record of infestation by a native tick (Acari: Ixodidae) on the Australian emu (*Dromaius novaehollandiae*) and a review of tick paralysis in Australian birds. *Experimental and Applied Acarology*, *73*(1), 103–107. Scopus. https://doi.org/10.1007/s10493-017-0168-0

Labruna, M. B., Leite, R. C., & Oliveira, P. R. de. (1997). Study of the Weight of Eggs from Six Ixodid Species from Brazil. *Memórias Do Instituto Oswaldo Cruz*, *92*, 205–207. https://doi.org/10.1590/S0074-02761997000200012

Lall, G. K., Darby, A. C., Nystedt, B., MacLeod, E. T., Bishop, R. P., & Welburn, S. C. (2010). Amplified fragment length polymorphism (AFLP) analysis of closely related wild and captive tsetse fly (*Glossina morsitans morsitans*) populations. *Parasites & Vectors*, *3*(1), 47. https://doi.org/10.1186/1756-3305-3-47

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Leo, S. S. T., Davis, C. S., & Sperling, F. A. H. (2012). Characterization of 14 microsatellite loci developed for *Dermacentor albipictus* and cross-species amplification in *D*. *andersoni* and *D*. *variabilis* (Acari: Ixodidae). *Conservation Genetics Resources*, *4*(2), 379–382. https://doi.org/10.1007/s12686-011-9553-x

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Linacre, A., & Tobe, S. (2013). *Wildlife DNA Analysis: Applications in Forensic Science*. John Wiley & Sons.

Lischer, H. E. L., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 474. https://doi.org/10.1186/s12859-017-1911-6

Loftis, A. D., Reeves, W. K., Szumlas, D. E., Abbassy, M. M., Helmy, I. M., Moriarity, J. R., & Dasch, G. A. (2006). Rickettsial agents in Egyptian ticks collected from domestic animals. *Experimental & Applied Acarology*, *40*(1), 67. https://doi.org/10.1007/s10493-006-9025-2

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science*, *3*, e104. https://doi.org/10.7717/peerj-cs.104

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(2047-217X-1–18). https://doi.org/10.1186/2047-217X-1-18

Mans, B. J., de Klerk, D., Pienaar, R., de Castro, M. H., & Latif, A. A. (2015). Next-generation sequencing as means to retrieve tick systematic markers, with the focus on *Nuttalliella namaqua* (Ixodoidea: Nuttalliellidae). *Ticks and Tick-Borne Diseases*, *6*(4), 450–462. https://doi.org/10.1016/j.ttbdis.2015.03.013

Marendy, D., Baker, K., Emery, D., Rolls, P., & Stutchbury, R. (2020). *Haemaphysalis longicornis*: The life-cycle on dogs and cattle, with confirmation of its vector status for *Theileria orientalis* in Australia. *Veterinary Parasitology: X*, *3*, 100022. https://doi.org/10.1016/j.vpoa.2019.100022

Masina, S., & Broady, K. W. (1999). Tick paralysis: Development of a vaccine. *International Journal for Parasitology*, *29*(4), 535–541. https://doi.org/10.1016/S0020-7519(99)00006-5

Masson, T., Fabre, M. L., Pidre, M. L., Niz, J. M., Berretta, M. F., Romanowski, V., & Ferrelli, M. L. (2021). Genomic diversity in a population of *Spodoptera frugiperda* nucleopolyhedrovirus. *Infection, Genetics and Evolution*, *90*, 104749. https://doi.org/10.1016/j.meegid.2021.104749

Mendoza-Roldan, J., Ribeiro, S. R., Castilho-Onofrio, V., Grazziotin, F. G., Rocha, B., Ferreto-Fiorillo, B., Pereira, J. S., Benelli, G., Otranto, D., & Barros-Battesti, D. M. (2020). Mites and ticks of

reptiles and amphibians in Brazil. *Acta Tropica*, *208*. Scopus. https://doi.org/10.1016/j.actatropica.2020.105515

Menzel, P., Ng, K. L., & Krogh, A. (2015). Kaiju: Fast and sensitive taxonomic classification for metagenomics. *BioRxiv*, 031229. https://doi.org/10.1101/031229

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Morin, P. A., Luikart, G., Wayne, R. K., & the SNP workshop group. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), 208–216. https://doi.org/10.1016/j.tree.2004.01.009

Motro, Y., & Moran-Gilad, J. (2017). Next-generation sequencing applications in clinical bacteriology. *Biomolecular Detection and Quantification*, *14*, 1–6. https://doi.org/10.1016/j.bdq.2017.10.002

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., & Gerstein, M. (2016). The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biology*, *17*(1), 53. https://doi.org/10.1186/s13059-016-0917-0

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., & Pati, A. (2015). Large-scale contamination of microbial isolate genomes by illumina Phix control. *Standards in Genomic Sciences*, *10*(APRIL2015). Scopus. https://doi.org/10.1186/1944-3277-10-18

Nadolny, R., Gaff, H., Carlsson, J., & Gauthier, D. (2015). Comparative population genetics of two invading ticks: Evidence of the ecological mechanisms underlying tick range expansions. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *35*, 153–162. https://doi.org/10.1016/j.meegid.2015.08.009

Narayanan, S. (1991). Applications of restriction fragment length polymorphism. *Annals of Clinical & Laboratory Science*, *21*(4), 291–296.

Navajas, M., & Fenton, B. (2000). The application of molecular markers in the study of diversity in acarology: A review. *Experimental & Applied Acarology; Dordrecht*, *24*(10/11), 751–774.

Neumann, L. G. (1899). Révision de la famille des ixodidés (3e mémoire). *Mémoires de La Société Zoologique de France*, *12*, 107–294. Scopus.

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 385. https://doi.org/10.1186/1471-2105-12-385

Pacheco, R. C., Echaide, I. E., Alves, R. N., Beletti, M. E., Nava, S., & Labruna, M. B. (2013). *Coxiella burnetii* in Ticks, Argentina. *Emerging Infectious Diseases*, *19*(2), 344–346. https://doi.org/10.3201/eid1902.120362

Parola, P., & Raoult, D. (2001). Ticks and tickborne bacterial diseases in humans: An emerging infectious threat. *Clinical Infectious Diseases*, *32*(6), 897–928. Scopus. https://doi.org/10.1086/319347

Parola, Philippe, Paddock, C. D., Socolovschi, C., Labruna, M. B., Mediannikov, O., Kernif, T., Abdad, M. Y., Stenos, J., Bitam, I., Fournier, P.-E., & Raoult, D. (2013). Update on Tick-Borne Rickettsioses around the World: A Geographic Approach. *Clinical Microbiology Reviews*, *26*(4), 657–702. https://doi.org/10.1128/CMR.00032-13

Pek, C. H., Cheong, C. S. J., Yap, Y. L., Doggett, S., Lim, T. C., Ong, W. C., & Lim, J. (2016). Rare Cause of Facial Palsy: Case Report of Tick Paralysis by *Ixodes Holocyclus* Imported by a Patient Travelling into Singapore from Australia. *The Journal of Emergency Medicine*, *51*(5), e109–e114. https://doi.org/10.1016/j.jemermed.2016.02.031

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., van Doorn, L.-J., Knetsch, C. W., & Figueiredo, C. (2019). Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.01277

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., &

Vizcaíno, J. A. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLOS Computational Biology*, *12*(7), e1004947. https://doi.org/10.1371/journal.pcbi.1004947

Piesman, J., & Stone, B. F. (1991). Vector competence of the Australian paralysis tick, *Ixodes holocyclus*, for the lyme disease spirochete Borrelia burgdorferi. *International Journal for Parasitology*, *21*(1), 109–111. https://doi.org/10.1016/0020-7519(91)90127-S

Playford, M. (2005). Review of research needs for cattle tick control—Phases I and II (AHW.054A). *Strategic Bovine Services*.

Poli, P., Lenoir, J., Plantard, O., Ehrmann, S., Røed, K. H., Leinaas, H. P., Panning, M., & Guiller, A. (2020). Strong genetic structure among populations of the tick *Ixodes ricinus* across its range. *Ticks and Tick-Borne Diseases*, *11*(6), 101509. https://doi.org/10.1016/j.ttbdis.2020.101509

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: Their purpose and place. *Human Molecular Genetics*, *27*(R2), R234–R241. https://doi.org/10.1093/hmg/ddy177

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(SUPPL. 1), D61–D65. Scopus. https://doi.org/10.1093/nar/gkl842

Radzijevskaja, J., Indriulytė, R., Paulauskas, A., Ambrasienė, D., & Turčinavičienė, J. (2005). Genetics Polymorphism Study of *Ixodes ricinus* L. Populations in Lithuania using RAPD Markers. *Acta Zoologica Lituanica*, *15*(4), 341–348. https://doi.org/10.1080/13921657.2005.10512699

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marçais, G., Pop, M., & Yorke, J. A. (2012). GAGE: A critical

evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. https://doi.org/10.1101/gr.131383.111

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, *109*(6), 365–371. https://doi.org/10.1007/s004120000089

Seker, A., Diri, B., Arslan, H., & Amasyalı, M. F. (2020). Open Source Software Development Challenges: A Systematic Literature Review on GitHub. *International Journal of Open Source Software and Processes (IJOSSP)*, *11*(4), 1–26. https://doi.org/10.4018/IJOSSP.2020100101

Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecology Letters*, *9*(5), 615–629. https://doi.org/10.1111/j.1461-0248.2006.00889.x

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Široký, P., Kubelová, M., Modrý, D., Erhart, J., Literák, I., Špitalská, E., & Kocianová, E. (2010). Tortoise tick *Hyalomma aegyptium* as long term carrier of Q fever agent *Coxiella burnetii*—Evidence from experimental infection. *Parasitology Research*, *107*(6), 1515–1520. https://doi.org/10.1007/s00436-010-2037-1

Šlapeta, J., Chandra, S., & Halliday, B. (2021). The “tropical lineage” of the brown dog tick *Rhipicephalus sanguineus* sensu lato identified as *Rhipicephalus linnaei* (Audouin, 1826). *International Journal for Parasitology*, *51*(6), 431–436. https://doi.org/10.1016/j.ijpara.2021.02.001

Sonenshine, D. E., & Roe, R. M. (2013). *Biology of Ticks Volume 1*. OUP USA.

Song, S., Shao, R., Atwell, R., Barker, S., & Vankan, D. (2011). Phylogenetic and phylogeographic relationships in *Ixodes holocyclus* and *Ixodes cornuatus* (Acari: Ixodidae) inferred from COX1 and ITS2 sequences. *International Journal for Parasitology*, *41*(8), 871–880. https://doi.org/10.1016/j.ijpara.2011.03.008

Sperling, J. L., Silva-Brandão, K. L., Brandão, M. M., Lloyd, V. K., Dang, S., Davis, C. S., Sperling, F. A. H., & Magor, K. E. (2017). Comparison of bacterial 16S rRNA variable regions for microbiome surveys of ticks. *Ticks and Tick-Borne Diseases*, *8*(4), 453–461. https://doi.org/10.1016/j.ttbdis.2017.02.002

Stajich, J. E., & Lapp, H. (2006). Open source tools and toolkits for bioinformatics: Significance, and where are we? *Briefings in Bioinformatics*, *7*(3), 287–296. https://doi.org/10.1093/bib/bbl026

Sunnucks, P. (2000). Efficient genetic markers for population biology. *Trends in Ecology & Evolution*, *15*(5), 199–203. https://doi.org/10.1016/S0169-5347(00)01825-5

Talbot, B., Leighton, P. A., & Kulkarni, M. A. (2020). Genetic Melting Pot in Blacklegged Ticks at the Northern Edge of their Expansion Front. *Journal of Heredity*, *111*(4), 371–378. https://doi.org/10.1093/jhered/esaa017

Tessler, M., Neumann, J. S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L. F. M., Segovia, B. T., Lansac-Toha, F. A., Lemke, M., DeSalle, R., Mason, C. E., & Brugler, M. R. (2017). Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, *7*(1), 6589. https://doi.org/10.1038/s41598-017-06665-3

Thankaswamy-Kosalai, S., Sen, P., & Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*, *109*(3), 186–191. https://doi.org/10.1016/j.ygeno.2017.03.001

Thomas, G. W. C., & Hahn, M. W. (2019). Referee: Reference Assembly Quality Scores. *Genome Biology and Evolution*, *11*(5), 1483–1486. https://doi.org/10.1093/gbe/evz088

Thrash, A., Hoffmann, F., & Perkins, A. (2020). Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics*, *21*(4), 249. https://doi.org/10.1186/s12859-020-3382-4

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Tufts, D. M., Sameroff, S., Tagliafierro, T., Jain, K., Oleynik, A., VanAcker, M. C., Diuk-Wasser, M. A., Lipkin, W. I., & Tokarz, R. (2020). A metagenomic examination of the pathobiome of the invasive tick species, *Haemaphysalis longicornis*, collected from a New York City borough, USA. *Ticks and Tick-Borne Diseases*, *11*(6), 101516. https://doi.org/10.1016/j.ttbdis.2020.101516

Unsworth, N. B., Stenos, J., Graves, S. R., Faa, A. G., Cox, G. E., Dyer, J. R., Boutlis, C. S., Lane, A. M., Shaw, M. D., Robson, J., & Nissen, M. D. (2007). Flinders Island Spotted Fever Rickettsioses Caused by "marmionii" Strain of *Rickettsia honei*, Eastern Australia. *Emerging Infectious Diseases*, *13*(4), 566–573. https://doi.org/10.3201/eid1304.060087

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—New capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115–e115. https://doi.org/10.1093/nar/gks596

Van Houtte, N., Van Oosten, A. R., Jordaens, K., Matthysen, E., Backeljau, T., & Heylen, D. J. A. (2013). Isolation and characterization of ten polymorphic microsatellite loci in *Ixodes arboricola*, and cross-amplification in three other *Ixodes* species. *Experimental and Applied Acarology*, *61*(3), 327–336. https://doi.org/10.1007/s10493-013-9702-x

van Nunen, S. (2015). Tick-induced allergies: Mammalian meat allergy, tick anaphylaxis and their significance. *Asia Pacific Allergy*, *5*(1), 3–16. https://doi.org/10.5415/apallergy.2015.5.1.3

Van Oosten, A. R., Heylen, D. J. A., Jordaens, K., Backeljau, T., & Matthysen, E. (2014). Population genetic structure of the tree-hole tick *Ixodes arboricola* (Acari: Ixodidae) at different spatial scales. *Heredity*, *113*(5), 408–415. https://doi.org/10.1038/hdy.2014.41

Veldsman, W. P., Ma, K. Y., Hui, J. H. L., Chan, T. F., Baeza, J. A., Qin, J., & Chu, K. H. (2021). Comparative genomics of the coconut crab and other decapod crustaceans: Exploring the molecular basis of terrestrial adaptation. *BMC Genomics*, *22*(1), 313. https://doi.org/10.1186/s12864-021-07636-9

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*. https://doi.org/10.12688/f1000research.15931.2

Wong, X. L., & Sebaratnam, D. F. (2018). Mammalian meat allergy. *International Journal of Dermatology*, *57*(12), 1433–1436. https://doi.org/10.1111/ijd.14208

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wu, X., Zhang, H., Chen, J., Shang, S., Yan, J., Chen, Y., Tang, X., & Zhang, H. (2017). Analysis and comparison of the wolf microbiome under different environmental factors using three different data of Next Generation Sequencing. *Scientific Reports*, *7*(1), 11332. https://doi.org/10.1038/s41598-017-11770-4

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

Yang, R., Murphy, C., Song, Y., Ng-Hublin, J., Estcourt, A., Hijjawi, N., Chalmers, R., Hadfield, S., Bath, A., Gordon, C., & Ryan, U. (2013). Specific and quantitative detection and identification of *Cryptosporidium hominis* and *C. parvum* in clinical and environmental samples. *Experimental Parasitology*, *135*(1), 142–147. https://doi.org/10.1016/j.exppara.2013.06.014

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Yessinou, R. E., Akpo, Y., Adoligbe, C., Adinci, J., Assogba, N., Koutinhouin, B., Karim, I. Y. A., & Farougou, S. (2016). Resistance of tick *Rhipicephalus microplus* to acaricides and control strategies. *Journal of Entomology and Zoology Studies*, 7.

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H., & Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets

and CAMI datasets. *BMC Bioinformatics*, *21*(1), 334. https://doi.org/10.1186/s12859-020-03667-3

6 - Appendix

6.1 - Microsatellite Retrieval Results of NCBI Datasets

Rhipicephalus linnaei

The program Krait v1.3.3 was able to detect a total of 1,081,011 STRs in the quality-controlled *R. linnaei* dataset with 464,512 being mono-nucleotide, 164,798 being di-nucleotide, 152,911 being tri-nucleotide, 285,539 being tetra-nucleotide, 11,302 being penta-nucleotide and 1,949 being hexa-nucleotide. Out of the 1,081,011 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 65,350 STR sequences. Of the 65,350 filtered STR sequences, 496 had primer sequences successfully designed using Primer3 (integrated into Krait), according to the parameters described in the methods section. Parsing this data for total STRs took 3h 38min 2s.

Krait was then used the detect STRs from MEGAHIT assembled FASTA contigs from *R. linnaei*. The program was able to detect 26,215 mono-nucleotide, 10,693 di-nucleotide, 10,425 tri-nucleotide, 15,251 tetra-nucleotide, 734 penta-nucleotide and 137 hexa-nucleotide STR sequences. Out of the 63,455 microsatellites discovered, the sequences with a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 2,753 STR sequences. Of the 2,753 filtered STR sequences, 1,236 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 4min 31s.

After this, Krait was used the detect STRs in the best-mapped data from bowtie2, which for the *R. linnaei* dataset was reads mapped to the *R. microplus* reference genome. The SAM to FASTA converted reads were input into Krait. The program was able to detect 189,191 mono-nucleotide, 105,399 di-nucleotide, 111,980 tri-nucleotide, 231,249 tetra-nucleotide, 7,922 penta-nucleotide and 1,451 hexa-nucleotide STR sequences. Out of the 647,192 microsatellites discovered, the sequences

that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 58,901 STR sequences. Of the 58,901 filtered STR sequences, 365 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 1h 15min 24s.

Finally, reads classified as Ixodidae by Kraken2 using the custom tick-db database were analysed using Krait. The program was able to detect 91,798 mono-nucleotide, 31,710 di-nucleotide, 30,896 tri-nucleotide, 53,817 tetra-nucleotide, 2,303 penta-nucleotide and 361 hexa-nucleotide STR sequences. Out of the 210,885 microsatellites discovered, the sequences with a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 12,375 STR sequences. Of the 12,375 filtered STR sequences, 477 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 32min 25s.

Rhipicephalus appendiculatus

The program Krait v1.3.3 was able to detect a total of 2,175,558 STRs in the quality-controlled *R. appendiculatus* dataset with 559,476 being mono-nucleotide, 209,938 being di-nucleotide, 126,142 being tri-nucleotide, 1269,222 being tetra-nucleotide, 8,020 being penta-nucleotide, and 2,760 being hexa-nucleotide. Out of the 2,175,558 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 236,251 STR sequences. Of the 236,251 filtered STR sequences, 10 had primer sequences successfully designed using Primer3 (integrated into Krait), according to the parameters described in the methods section. Parsing this data for total STRs took 2h 9min 52s.

Krait was then used the detect STRs in MEGAHIT assembled FASTA contigs from *R. appendiculatus*. The program was able to detect 21,502 mono-nucleotide, 5,928 di-nucleotide, 5,676 tri-nucleotide,

20,321 tetra-nucleotide, 302 penta-nucleotide and 99 hexa-nucleotide STR sequences. Out of the 53,828 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 3,778 STR sequences. Of the 3778 filtered STR sequences, 1,364 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 4min 35s.

After this, Krait was used the detect STRs in the best-mapped data from bowtie2, which for the *R. appendiculatus* dataset was reads mapped to the *R. microplus* reference genome. The SAM to FASTA converted reads were input into Krait. The program was able to detect 559,476 mono-nucleotide, 209,938 di-nucleotide, 126,142 tri-nucleotide, 1269,222 tetra-nucleotide, 8,020 penta-nucleotide and 2,760 hexa-nucleotide STR sequences. Out of the 2,175,558 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 236,251 STR sequences. Of the 236,250 filtered STR sequences, 10 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 1h 45min 5s.

Reads classified as Ixodidae by Kraken2 using the custom tick-db database were analysed using Krait. The program was able to detect 279,991 mono-nucleotide, 93,629 di-nucleotide, 90,978 tri-nucleotide, 765,035 tetra-nucleotide, 4,616 penta-nucleotide and 1,669 hexa-nucleotide STR sequences. Out of the 1,235,918 microsatellites discovered, the sequences with a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 139,720 STR sequences. Of the 139,720 filtered STR sequences, 13 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 1h 2min 8s.

Dermacentor variabilis

The program Krait v1.3.3 was able to detect a total of 7,813,899 STRs in the quality-controlled *D. variabilis* dataset with 1,611,341 being mono-nucleotide, 1,781,774 being di-nucleotide, 1,680,173 being tri-nucleotide, 2,430,138 being tetra-nucleotide, 289,386 being penta-nucleotide, and 21,087 being hexa-nucleotide. Out of the 7,813,899 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 245,088 STR sequences. Of the 245,088 filtered STR sequences, two had primer sequences successfully designed using Primer3 (integrated into Krait), according to the parameters described in the methods section. Parsing this data for total STRs took 45h 31m 20s.

Krait was then used the detect STRs from MEGAHIT assembled FASTA contigs from *D. variabilis*. The program was able to detect 83,837 mono-nucleotide, 90,089 di-nucleotide, 64,624 tri-nucleotide, 81,033 tetra-nucleotide, 3,434 penta-nucleotide and 845 hexa-nucleotide STR sequences. Out of the 323,862 microsatellites discovered, the sequences with a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 22,634 STR sequences. Of the 22,634 filtered STR sequences, no primer sequences were successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 31min 39s.

After this, Krait was used the detect STRs in the best-mapped data from bowtie2, which for the *D. variabilis* dataset was reads mapped to the *D. silvarum* reference genome. The SAM to FASTA converted reads were input into Krait. The program was able to detect 585,600 mono-nucleotide, 1,216,755 di-nucleotide, 1,232,364 tri-nucleotide, 1,964,465 tetra-nucleotide, 270,654 penta-nucleotide and 15,103 hexa-nucleotide STR sequences. Out of the 5,284,941 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 227,843 STR sequences. Of the 227,843 filtered STR sequences, two

had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 19h 32min 28s.

Reads classified as Ixodidae by Kraken2 using the custom tick-db database were analysed using Krait. The program was able to detect 324,339 mono-nucleotide, 467,972 di-nucleotide, 796,168 tri-nucleotide, 1,109,994 tetra-nucleotide, 201,348 penta-nucleotide and 6,595 hexa-nucleotide STR sequences. Out of the 2,906,416 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 97,299 STR sequences. Of the 97,299 filtered STR sequences, no had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 17h 17min 44s.

Tick Virome

The program Krait v1.3.3 was able to detect a total of 14,082 STRs in the quality-controlled Tick Virome dataset with 2,571 mono-nucleotide, 3,632 di-nucleotide, 4,571 tri-nucleotide, 2,942 tetra-nucleotide, 268 penta-nucleotide and 98 hexa-nucleotide. Out of the 14,082 microsatellites discovered, the sequences with a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 269 STR sequences. Of the 269 filtered STR sequences, none had primer sequences successfully designed using Primer3 (integrated into Krait), according to the parameters described in the methods section. Parsing this data for total STRs took 2min 32s.

After this, Krait was used the detect STRs in the best-mapped data from bowtie2, which for the Tick Virome dataset was reads mapped to the *H. longicornis* reference genome. The SAM to FASTA converted reads were input into Krait. The program was able to detect 2,362 mono-nucleotide, 3,591 di-nucleotide, 4,552 tri-nucleotide, 2,921 tetra-nucleotide, 260 penta-nucleotide and 91 hexa-nucleotide STR sequences. Out of the 13,777 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 269 STR

sequences. Of the 269 filtered STR sequences, none had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 2min 10s.

Finally, reads classified as Ixodidae by Kraken2 using the custom tick-db database were analysed using Krait. The program was able to detect 1,417 mono-nucleotide, 1,508 di-nucleotide, 1,050 tri-nucleotide, 1,922 tetra-nucleotide, 222 penta-nucleotide and 17 hexa-nucleotide STR sequences. Out of the 6,136 microsatellites discovered, the sequences that had a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was 178 STR sequences. Of the 178 filtered STR sequences, 92 had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 1min 55s.

After this, Krait was used the detect STRs in the best-mapped data from MEGAHIT assembled FASTA contigs from the Tick Virome dataset. The program was able to detect 70 mono-nucleotide, 82 di-nucleotide, 80 tri-nucleotide, 117 tetra-nucleotide, 11 penta-nucleotide and 1 hexa-nucleotide STR sequences. Out of the 361 microsatellites discovered, the sequences with a motif of three or more, a start flanking region of 20 or more and a length of 50 or more was four STR sequences. Of the four filtered STR sequences, two had primer sequences successfully designed using Primer3, according to the parameters described in the methods section. Parsing this data for total STRs took 7s.