



University of  
**Salford**  
MANCHESTER

# THESIS

Genomic Characterisation of Novel Veterinary  
Pathogens: *Anaplasma* & *Bartonella* species

**Sean Brierley**

**Oct 2024**

# ABSTRACT

**Background:** *Bartonella* sp. and *Anaplasma phagocytophilum* (*Ap*) are vector-borne bacterial pathogens with significant veterinary and public health implications. While *Bartonella* species persist in the bloodstream of various mammals causing long term bacteraemia, *Ap* is an intracellular pathogen causing granulocytic anaplasmosis. Despite their importance, genomic data on novel *Bartonella* species and UK *Ap* strains remains limited. Additionally, the low abundance and intracellular nature of *Ap* complicate direct sequencing from host tissues. Expanding genomic resources and refining enrichment methods are essential for improving pathogen characterisation and understanding host-pathogen interactions.

**Objectives:** Characterise a novel species of *Bartonella*, generate the first complete genome representations of *Anaplasma phagocytophilum* (*Ap*) isolated in the UK and develop optimised enrichment methodologies for high-resolution sequencing of *Ap* directly from infected host tissue.

**Methods:** Three *Bartonella* strains isolated from field voles (*Microtus agrestis*) and seven *Ap* strains isolated from domestic ruminants were sequenced using Illumina short-read and Oxford Nanopore long-read systems. Genomic analyses included phylogenetic reconstruction based on concatenated core gene alignments, pangenomic profiling, and average nucleotide identity calculations. Enrichment strategies encompassing differential lysis (Molzym), CpG methylation depletion (NEB), biotinylated RNA bait capture (Agilent SureSelect), and adaptive sampling (ONT) were systematically evaluated on roe deer spleen samples infected with *Ap*. Alignment files were investigated to assess genome coverage and identify capture biases. An optimised approach was applied to the spleen of an *Ap*-infected common shrew (*Sorex araneus*) with the aim of characterising the currently uncultured, genetically divergent small mammal-associated (ecotype III) strain of the species.

**Results:** Whole genome analyses identified the three *Bartonella* strains as a novel lineage 3 species, proposed as *Bartonella bennettii* most notably containing a chromosomally integrated *vbh/TraG* type IV secretion system of plasmid origin.

Phylogenetic analysis of UK *Ap* isolates placed them within the European ecotype I cluster, while revealing potential subdivisions. The pangenome identified core and accessory genes, with ANI values suggesting species boundaries within *Ap*.

Enrichment protocols combining Monarch HMW DNA extraction and NEB microbiome depletion yielded optimal pathogen representation. Gap analysis highlighted capture biases and the potential of the technology to capture complete *Ap* genomes, especially in the context of long read systems. The small mammal-associated ecotype III strain was partially captured with non-specific ecotype I baits identifying the limits of the capture technology. Linkage analysis of *groEL* genes supported existing ecotype classifications, whereas whole genome phylogenetics indicated potential reclassification into four epidemiologically separated species in a global context.

**Conclusions:** *B. bennettii* was characterised through genomic analyses, providing insights into the diversity and evolution of virulence factors in the *Bartonella* genus. Additionally, the first complete *Ap* genomes from the UK were generated, providing insights into genomic diversity and phylogenetic relationships. Optimised enrichment strategies were developed for high-resolution metagenomic sequencing, overcoming challenges posed by low bacterial loads and complex metagenomic samples. Whole genome analysis suggests the European ecotypes are representative of global *Ap* diversity, with ANI supporting the existence of four epidemiologically separate species within *Ap*. Continued genomic characterisation is crucial for understanding the drivers of host specificity, zoonotic potential, and epidemiological dynamics within these diverse blood-borne parasites.

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the many individuals and organisations who have supported me throughout the completion of this thesis.

First and foremost, I extend my sincere thanks to my professors Richard Birtles, Ian Goodhead, and Dr Kevin Bown. Their guidance, expertise, and encouragement have been invaluable throughout my research.

I would also like to thank my peers in the lab for their constant support and endless advice. Their camaraderie and insights have greatly enriched my research experience.

Special thanks go to the SEM shared research facility and Alison Beckett for her crucial assistance in capturing *B. bennettii*. Your expertise and support were essential to the characterisation of this novel bacterium.

I am also grateful to the staff at Agilent for their help in developing and troubleshooting the bait capture technology. Your technical support and dedication have been instrumental.

Furthermore, I would like to thank Forestry England and the abattoir and farms that contributed to the collection of spleens and blood samples.

Finally, I would like to specially mention my family and my girlfriend. Your unwavering support, understanding, and encouragement have been my greatest strength throughout this process. Thank you for being my foundation.

To all those mentioned and to anyone I may have inadvertently left out, your support has been deeply appreciated.



# TABLE OF CONTENTS

<b>CHAPTER 1</b>	<b>2</b>
<b>1.0: Introduction</b>	<b>2</b>
1.1: Classification & Characteristics of Bartonellae	5
1.2: Hosts, Reservoirs & Vectors	11
1.3: Molecular Basis of Bartonellae Parasitism	14
1.4: <i>Bartonella</i> Genome Evolution	17
1.5: Classification & Identification of <i>Anaplasma phagocytophilum</i>	21
1.6: Hosts, Reservoirs and Vectors	24
1.7: Diversity & Strategies of <i>Anaplasma phagocytophilum</i>	29
1.8: Conclusion	37
1.9: Aims & Objectives	38
<b>CHAPTER 2</b>	<b>40</b>
<b>2.0: Characterisation of a Novel Small Mammal-Associated <i>Bartonella</i> Species</b>	<b>40</b>
2.1 Introduction	40
2.2: Methods	45
2.3: Results	49
2.4: Discussion	64
<b>CHAPTER 3</b>	<b>71</b>
<b>3.0: Exploring the Genomic Diversity of UK <i>Anaplasma phagocytophilum</i> Isolates.</b>	<b>71</b>
3.1: Introduction	71
3.2: Methods	74

3.3: Results	79
3.4: Discussion	89
<b>CHAPTER 4</b>	<b>96</b>
<b>4.0: Development &amp; Optimisation of Enrichment Protocols for High-Resolution Sequencing of <i>Anaplasma phagocytophilum</i> Genomes from Blood &amp; Tissue.</b>	<b>96</b>
4.1: Introduction	96
4.2: Methods	100
4.3: Results	104
4.4: Discussion	115
<b>CHAPTER 5</b>	<b>121</b>
<b>5.0: Ecotype Resolution in <i>Anaplasma phagocytophilum</i> Applying Enrichment Strategies to a Global Context</b>	<b>121</b>
5.1: Introduction	121
5.2: Methods	124
5.3: Results	127
5.4: Discussion	131
<b>CHAPTER 6</b>	<b>136</b>
<b>6.0: Unravelling the Diversity &amp; Mysteries of Blood-Borne Pathogens in the UK &amp; Beyond – A Comprehensive Discussion.</b>	<b>136</b>
6.1: Introduction	136
6.2: Key Findings	137
6.3: Contextualising Research Findings	140
6.4: Future Directions	144
6.5: Conclusion	146

# LIST OF TABLES

Table 1: A collection of the best genome representations of all validly published species of <i>Bartonella</i> that have available data; retrieved from NCBI GenBank	18
Table 2: A collection of all publicly available <i>Anaplasma phagocytophilum</i> strains sequenced across the world. Source: NCBI GenBank, accessions can be used to access all strains in column 1.	34
Table 3: <i>Bartonella</i> species currently without standing in nomenclature.	42
Table 4: A collection of partially characterised lineage 3 strains from the <i>Bartonella</i> genus. NCBI GenBank accession numbers are available in column 1.	43
Table 5: Genome statistics for three strains of <i>Bartonella bennettii</i> sp. nov. (C271, D105, J177) and <i>B. hiexiaziensis</i> strain RE21. CDS annotations were generated with RASTtk.	53
Table 6: ANI matrix of proposed <i>Anaplasma phagocytophilum</i> ecotype representatives calculated with PyANI.	128
Table 7: Total number of CDS annotated on Seven UK derived isolates of <i>A. phagocytophilum</i> (Harris, Perth, ZW144, ZW122, ZW129, OS, FG) and three strains of <i>Bartonella bennettii</i> sp. nov. (C271, J117, D105) and <i>B. heixiaziensis</i> (RE21).	178
Table 8: Assembly statistics & sequencing data generated for seven UK derived isolates of <i>A. phagocytophilum</i> .	178

# LIST OF FIGURES

- Figure 1: The bartonellae lineages delineated using core genome concatenations and maximum likelihood inference (Engel et al., 2011). 7
- Figure 2: Scanning electron microscopy of *Bartonella kosoyi* Tel Aviv strain, showing rod shaped bacteria interconnected by polar pili with a 500nm scale bar for reference (Gutierrez et al., 2020). 9
- Figure 3: The common strategy of Bartonellae to establish infection in reservoir hosts. (a) Bartonellae colonise the dermis (b) *Bartonella* co-opt migratory cells to the blood-seeding niche (c) Bartonellae establish infection in endothelial cells (d) Bartonellae are seeded into the blood stream where they invade the erythrocytes and other endothelial cells (e) Bartonellae replicate with host erythrocytes (f) inducing apoptosis (g) blood meals of the infected host transmit *Bartonella* back into arthropod vectors (Harms & Dehio, 2012). 14
- Figure 4: Phylogenetic tree of the *Bartonella* genus based on conserved core genome homology. Vector, reservoir zoonotic capabilities and virulence factors present in each species is indicated (Wagner & Dehio, 2019). 19
- Figure 5: A peripheral blood smear from a cat infected with *Anaplasma phagocytophilum* stained using Wright-Giemsa stain. The larger arrow points to a morula within the cytoplasm of a neutrophil. The smaller arrow points to a Dohle body for comparison (Headquarters, 2016). 22
- Figure 6: Phylogenetic tree based on seven concatenated housekeeping gene sequences (2877 bp) of *A. phagocytophilum*. Phylogenetic analysis was performed with the maximum likelihood method based on the Tamura-Nei model in MEGA 6.0 with 1000 bootstrap replicates (Huhn et al., 2014). 31
- Figure 7: The *msp2* expression loci of *Anaplasma marginale* and *Anaplasma phagocytophilum*. The black boxes represent potential promoter regions for expression of *msp2* and upstream genes in the loci (Barbet et al., 2005). 35
- Figure 8: *Msp2* expression site variability in three US and two European strains of *Anaplasma phagocytophilum*. A similarity comparison where a score of 1.0 on the Y axis equals identical, with values closer to 0 representing increasing variation (Barbet et al., 2006). 36
- Figure 9: Scanning electron microscopy (SEM) of *B. bennettii* strain C271 performed at the Liverpool SEM shared research facility. Three images show population structure (15000x), edge of colony (80000x) and an isolated cell displaying polar lophotrichous flagella (80000x). 50
- Figure 10: A Circos plot of the *Bartonella sp.* strain C271. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). 51
- Figure 11: A Circos plot of the *Bartonella sp.* strain D105. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). 51
- Figure 12: A Circos plot of the *Bartonella sp.* strain J117. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). 52
- Figure 13: A Circos plot of the *Bartonella sp.* strain RE21. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). 52

- Figure 14: Phylogenetic tree of Lineage 3 *Bartonella* strains and *B. rochalimae* (Br) and *B. clarridgeiae* (Bc). Strains D105, C271, J117 and AR15-3 are proposed as a novel lineage 3 species *B. bennettii* (Bb). Tree was generated with RAxML using 619,662 amino acid alignments of 500 core single copy genes for each species, bootstrapped with 100 replicates. Strains from this study are marked with \* 53
- Figure 15: Core genome tree of 48 *Bartonella* genomes. 100 gene maximum-likelihood tree generated from an alignment of 39671 amino acids and bootstrapped with 100 replicates. The presence and absence of key virulence factors is indicated in columns (P) indicates that the *vbh/TraG* type 4 secretion system is present on a plasmid. 54
- Figure 16: Average nucleotide identity heatmap of 48 *Bartonella* genomes calculated with pyANI. Values available in the appendix. 57
- Figure 17: Circos plot illustrating synteny between two *Bartonella* genomes RE21 (blue), C271 (purple) from species *B. hiexiaziensis* and *B. bennettii* respectively. (Orange highlight) likely bacterial conjugation event of a plasmid containing the *vbh/TraG* T4SS. (Yellow highlights) *virB/D4* T4SS locations. 57
- Figure 18: Synteny comparison generated in Clinker of the small *virB/D4* type 4 secretion system in lineage 3 species of *Bartonella*. (Br) *Bartonella rochalimae*, (Bc) *Bartonella clarridgeiae*, (C271, D105, J117, AR-15-3) *B. bennettii*. 59
- Figure 19: Synteny comparison generated in Clinker of the large *virB/D4* type 4 secretion system in lineage 3 species of *Bartonella*. (Br) *Bartonella rochalimae*, (Bc) *Bartonella clarridgeiae*, (C271, D105, J117, AR-15-3) *B. bennettii*. 59
- Figure 20: Synteny comparison generated in Clinker of the *virB/D4* T4SS in lineage 4 *Bartonella* species, (Bh) *Bartonella henselae*, (Bq) *Bartonella quintana*, (Bt) *Bartonella tribocorum*, (Bg) *Bartonella grahamii*, (RE21) *Bartonella hiexiaziensis*. 60
- Figure 21: Synteny comparison of the *vbh/TraG* T4SS in strains of *Bartonella bennettii* sp. nov. generated in Clinker. The System is absent in strain AR15-3. 60
- Figure 22: Pangenome analysis of the *Bartonella* genus generated with Anvi'o displaying 48 species. The layers represent individual genomes organised by phylogenetic relationships based on 295 single copy core genes. Colours represent distinct *Bartonella* lineages (L1 -L4) which are orange, red, purple and blue respectively. Genome completeness, redundancy, GC content, length, genes per kbp singletons, and total number of gene clusters are indicated on the right. On the bottom NCBI COG20 functions (Green line indicates a function was successfully assigned to a gene cluster), number of genomes that contain a specific gene cluster and number of contributing genomes is indicated. In addition to this, the core, accessory and singleton gene clusters are highlighted with purple, green and orange respectively. 61
- Figure 23: The four ecotypes of *Anaplasma phagocytophilum* detected in European wildlife, their respective host tropisms and vectors. 72
- Figure 24: A Circos plot of the *Anaplasma phagocytophilum* Harris strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 79
- Figure 25: A Circos plot of the *Anaplasma phagocytophilum* OS strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 80
- Figure 26: A Circos plot of the *Anaplasma phagocytophilum* ZW129 strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 81
- Figure 27: A Circos plot of the *Anaplasma phagocytophilum* Perth strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology

to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 82

Figure 28: A Circos plot of the *Anaplasma phagocytophilum* ZW122 strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 83

Figure 29: A Circos plot of the *Anaplasma phagocytophilum* ZW144 strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 84

Figure 30: A Circos plot of the *Anaplasma phagocytophilum* FG strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. 85

Figure 31: Maximum-likelihood phylogenetic tree generated from a concatenated alignment of amino acid sequences of 500 single copy core genes within the *Anaplasma phagocytophilum* (*Ap*) species, generated with RAxML using the DUMMY2 protein model. Scale bar represents 0.007 amino acid substitutions per site and colour delineate known clusters/ variants. Dark blue represents strains of European origin, determined to be ecotype I. Light blue represents UK derived ecotype I strains, yellow represents the North American human active variants of *Ap* (*Ap*-HA) and red represents the North American variant 1 strains (*Ap*-V1). 86

Figure 32: Average nucleotide identity heatmap of 40 *Anaplasma phagocytophilum* genomes calculated with pyANI. 87

Figure 33: Pangenome analysis of *Anaplasma phagocytophilum* generated with Anvi'o displaying 40 strains. The layers represent individual genomes organised by phylogenetic relationships based on 500 single copy core genes. Colours represent distinct *Ap* clusters. Genome completeness, redundancy, GC content, length, genes per kbp singletons, and total number of gene clusters are indicated on the right. On the bottom NCBI COG20 functions, number of genomes that contain a specific gene cluster and number of contributing genomes is indicated. In addition to this, the core, accessory and singleton gene clusters are highlighted with red, blue and green respectively. 88

Figure 34: Map of Cumbria, UK with the locations of farms which provided sheep for slaughter at the abattoir on the 8<sup>th</sup> of November 2022. The location of Grizedale Forest is also indicated where deer spleens were collected from the 1<sup>st</sup> of November 2022 to 24<sup>th</sup> March 2023. 101

Figure 35: Relative infection intensity of *Anaplasma phagocytophilum* in red deer, roe deer and sheep collected in Cumbria, UK between November 2022 and March 2023. Measured with a Bio-Rad Opticon using a real-time PCR assay targeting a 77bp fragment of *msh2* 105

Figure 36: Benchmark sequencing efficiency achieved on the PromethION P2 solo utilising the NEB Monarch HMW DNA extraction kit for tissue on infected roe deer spleens (GRD08, GRD09, GRD17). Cycle threshold (Ct) is indicated for all three samples ranging from 21.2 - 19.0. Relative population sizes of host, other and *Anaplasma phagocytophilum* (*Ap*) are estimated using Kraken 2 read classification data generated against a custom database 106

Figure 37: The relationship between cycle threshold (infection intensity) as measured by a real-time PCR assay targeting a 77bp fragment of the *MSP2* gene and the percentage of reads identified as *Anaplasma phagocytophilum* when sequenced with the PromethION P2 solo with adaptive sampling active and enriching for the Harris genome. 107

Figure 38: The relationship between cycle threshold (infection intensity) as measured by a real-time PCR assay targeting a 77bp fragment of the *MSP2* gene and the percentage of reads identified as *Anaplasma phagocytophilum* after bait capture and sequencing with the Agilent SureSelect platform and Illumina MiSeq V3 respectively. 108

**Figure 39:** The percentage share of *Anaplasma phagocytophilum* (Ap) reads in a sequencing dataset of 3 samples (GRD08, GRD09, GRD17) subjected to four different enrichment techniques: (1) Monarch HMW DNA extraction and SureSelect bait capture; (2) Monarch HMW DNA extraction, NEB microbiome enrichment, SureSelect bait capture; (3) Molzym Complete5 extraction and SureSelect bait capture; (4) Molzym Complete5 extraction, NEB microbiome enrichment and SureSelect bait capture. Green bars (Ap) represent reads classified as *Anaplasma phagocytophilum* (Ap), dark grey bars (Other) represent reads classified as non-target microbes, light grey bars (Host) represent reads classified as roe deer. Classification of reads was performed with Kraken 2 using a custom database. 110

**Figure 40:** The percentage share of *Anaplasma phagocytophilum* (Ap) reads in a sequencing dataset of 3 samples (GRD08, GRD09, GRD17) subjected to four different enrichment techniques with (AS, green) or without (None, red) adaptive sampling: (1) Monarch HMW DNA extraction (2) Monarch HMW DNA extraction and NEB microbiome enrichment; (3) Molzym Complete5 extraction; (4) Molzym Complete5 extraction and NEB microbiome enrichment. Classification of reads was performed with Kraken 2 using a custom database. 110

**Figure 41:** Coverage statistics for reads generated for a roe deer derived *Anaplasma phagocytophilum* (Ap) strain GRD08 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository. 112

**Figure 42:** Coverage statistics for reads generated for a roe deer derived *Anaplasma phagocytophilum* (Ap) strain GRD09 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository. 113

**Figure 43:** Coverage statistics for reads generated for a roe deer derived *Anaplasma phagocytophilum* (Ap) strain GRD17 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository. 114

**Figure 44:** Linkage disequilibrium analysis of 46 *Anaplasma phagocytophilum* strains, displaying 23 haplotypes across four main ecotypes. *GroEL* fragments (648bp) aligned with Muscle integrated into MEGA11. MEGA11 was used to generate a maximum-likelihood tree bootstrapped with 100 replicates. Alignment and tree data was inputted into the Haploviewer software package. 128

**Figure 45:** Coverage statistics for reads generated for a common shrew derived *Anaplasma phagocytophilum* (Ap) strain CS5 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository. 129

**Figure 46:** Maximum-likelihood phylogenetic tree generated from a concatenated alignment of amino acid sequences of 500 single copy core genes within the *Anaplasma phagocytophilum* (Ap) species, generated with RAxML using the DUMMY2 protein model. Scale bar represents 0.01 amino acid substitutions per site and the colours delineate known clusters/ variants. Dark blue represents strains of European origin, determined to be ecotype I. Light blue represents UK derived ecotype I strains, yellow represents the North American human active variants of Ap (Ap-HA) and red represents the North American variant 1 strains (Ap-V1). 130

# LIST OF ABBREVIATIONS

<b>Ap:</b> <i>Anaplasma phagocytophilum</i>	<b>rtPCR:</b> Real-time polymerase chain reaction
<b>AS:</b> Adaptive Sampling	<b>MLST:</b> Multi-locus sequence typing
<b>BaGTA:</b> <i>Bartonella</i> gene transfer agent	<b>TEM:</b> Transmission electron microscopy
<b>Beps:</b> <i>Bartonella</i> effector proteins	<b>SEM:</b> Scanning electron microscopy
<b>CT:</b> Cycle Threshold	<b>UK:</b> United Kingdom
<b>GA:</b> Granulocytic anaplasmosis	<b>USA:</b> United States of America
<b>HGA:</b> Human granulocytic anaplasmosis	<b>CFAs:</b> CAMP-like factor autotransporters
<b>HMW:</b> High molecular weight	<b>LPS:</b> Lipopolysaccharide
<b>L1-L4:</b> Lineage 1 – 4	<b>KDO:</b> Keto-deoxyoctulosonic acid
<b>ONT:</b> Oxford Nanopore Technologies	<b>ITS:</b> Internal transcribed spacer
<b>PCR:</b> Polymerase chain reaction	<b>BID:</b> <i>Bartonella</i> intracellular delivery
<b>RcGTA:</b> <i>Rhodobacter capsulatus</i> gene transfer agent	<b>Br:</b> <i>Bartonella rochalimae</i>
<b>T4CP:</b> Type 4 coupling protein	<b>Bc:</b> <i>Bartonella clarridgeiae</i>
<b>TEM:</b> Transmission electron microscope	<b>Bb:</b> <i>Bartonella bennettii</i>
<b>T4SS:</b> Type 4 secretion system	<b>PTMs:</b> Post translational modifications
<b>TBF:</b> Tick-borne fever	<b>WGS:</b> Whole genome sequencing
<b>TP:</b> Tick pyaemia	<b>ANI:</b> Average nucleotide identity
<b>WGS:</b> Whole genome sequencing	<b>CDS:</b> Coding sequence
<b>NEB:</b> New England Biolabs	<b>PFA:</b> Paraformaldehyde
<b>COG:</b> Cluster of orthologous genes	<b>GA:</b> Glutaraldehyde
<b>MAGs:</b> Metagenomic assembled genomes	



# CHAPTER 1

## 1.0: Introduction

Since microorganisms were first recognised, researchers have devised methods of systematically grouping and distinguishing strains and species in an evolutionary and phylogenetic context (Clark, 1985). Early attempts at characterising bacteria were largely founded in biochemical, physiological and phenotypic differences such as size and shape (Emerson, Agulto, Liu, & Liu, 2008). There is huge biochemical variation in bacteria that influences both their metabolism and cell structure (Smits, Kuipers, & Veening, 2006). These differences have proved to be particularly useful in characterising certain groups, however, many bacteria simply cannot be conclusively distinguished using these traditional methods (Janda, & Abbott, 2002). As a result, microbiologists strived to develop new tools to improve the accuracy of taxonomic classifications. A paper by Schildkraut et al., 1961 describes the first attempt at microbial classification based on single stranded DNA. Schildkraut used *in vitro* hybridisation of DNA molecules to determine taxonomic relationships between microorganisms, where the conjugation of DNA molecules indicates genetic and taxonomic relatedness, a major breakthrough at the time. Perhaps most importantly, the generation of this type of data paved the way for our modern polyphasic classification system. The term polyphasic in a taxonomic context was coined by Colwell in 1970 and refers to the integration of chemotypic, phenotypic and genotypic data to generate natural, authentic and reliable groupings of microorganisms (Colwell, 1970). DNA hybridisation was quickly superseded by more modern genotypic typing methods such as 16S rRNA sequencing and molecular fingerprinting techniques and more recently whole genome sequencing. The molecular revolution has transformed all aspects of biology (Furrie, 2006; Carrico, Sabat, Friedrich, & Ramirez, 2013) One major early discovery brought about by the molecular comparison of evolutionarily conserved ribosomal genes, proposed that bacteria were in fact made up of two separate domains, the bacteria and the archaea, each equally as distinct from each other as they are from eukaryotes

demonstrating the limited discriminatory capacity of chemotypic and phenotypic characterisation techniques (Woese, 1987).

The classification, characterisation and identification of bacterial species and strains remains of vital importance in many areas of modern civilisation. This includes public health (Oliver, 2000), clinical diagnosis (Sibley, Peirano, & Church), environment monitoring (Sumampouw, & Risjani, 2014), food safety (O'Sullivan, Ross, & Hill, 2002), and identification of potential biological threat agents (Iqbal et al., 2000). Advancements made in molecular biology have led to newer fields of research such as genomics and proteomics, the study of the genome and the proteins produced by the genome respectively. These fields offer an enticing alternative to traditional microbiological techniques for characterising and identifying bacteria as they provide taxonomically relevant, multidimensional, molecular data, in a timely manner (Emerson et al., 2008). Next generation sequencers have enabled a broader and deeper look into the world of microorganisms through increasingly accessible comprehensive characterisation (Cao, Fanning, Proos, Jordan, & Srikumar, 2017). The reduced labour and cost of sequencing genomes has enabled both small and large research groups to generate draft sequences, thus, modern bacterial classification has become dominated by genomic studies (Kulski, 2016; Paul, Raj, Murali, & Satyamoorthy, 2020). Whole genome comparative genomics is the gold standard for classification of bacteria with many bacterial species being reclassified after genome-genome comparisons (Stropko, Pipes, & Newman, 2014; Lawson, Citron, Tyrrell, & Finegold, 2016). It is now a formal requirement to produce a complete or near complete draft genome sequence to define novel species of bacteria. Despite this, even with a vastly reduced cost and time to sequence, many genomes are yet to be fully sequenced, which means researchers must rely upon previously established molecular marker-based identification methods to investigate phylogenetic and evolutionary relationships. Over the decades, many different molecular markers have been used for phylogenetic classification. Characteristics of these markers are vast, ranging from highly conserved regions, variable regions, specific genes, and species-specific repeat elements (Paul et al, 2020). Of these markers, 16S rRNA has been extensively used due to a mosaic structure of highly conserved and hypervariable regions within its ~1500bp (Liu, Li, Khan, & Zhu, 2012; Rosselli et al, 2016). However, for several closely related species or strains,

comparison of the 16S rRNA sequence has rendered inconclusive results due to a high degree of similarity (Paul et al, 2020). Therefore, when whole genome comparisons are not available and 16S rRNA comparisons are inadequate, species-specific molecular markers are utilised, which provide a greater degree of variation to distinguish these closely related species. These markers are usually in the form of a specific gene, for example, the *Bartonella* genus is commonly characterised using partial ~300bp sequences of *gltA*, the citrate synthase gene (Norman, Regnery, Jameson, Greene, & Krause, 1995; Gutierrez et al, 2020). Equally, species-specific molecular markers are used to distinguish microbes at a strain level, for example, *Anaplasma phagocytophilum* (Ap), a complex and highly diverse veterinary pathogen can be delineated into ecotypes based on the *groEL* gene (Jahfari et al., 2014).

*Anaplasma* and *Bartonella*, two distinct genera of Alphaproteobacteria, share a noteworthy link through their common association with arthropod vectors (Dantas-Torres & Otranto, 2017; Billeter et al., 2008). These Gram-negative intracellular bacteria share an ability to infect a wide range of mammalian hosts, including humans, and each have constituent species recognised as emerging pathogens that have far-reaching implications for both animal and public health (Rar et al., 2021; Krügel et al., 2022). The threat to agricultural productivity and economic stability is bolstered by a zoonotic capacity that underscores the interconnectedness of veterinary and public health domains. The study of these pathogens is crucial for safeguarding not only the health and welfare of animals but also for maintaining biosecurity and mitigating the risks of emerging infectious diseases. Beyond their shared reliance on arthropod vectors, *Anaplasma* and *Bartonella* exhibit intriguing parallels in their pathogenic mechanisms and their ability to manipulate host cells. Both genera are adept at evading the hosts immune response, establishing persistent infections that contribute to the chronic nature of the infections they cause (Rar et al., 2021; Krügel et al., 2022). Both *Anaplasma* and *Bartonella* share a predilection for infecting the blood. This convergence in cellular tropism highlights the sophisticated strategies employed by constituent species that enable them to colonise and thrive within their hosts. Interestingly, co-infection of *Anaplasma* and *Bartonella* is a distinct possibility in many hosts, including roe deer and raises questions about potential synergistic effects on pathogenesis and inter-specific interactions (Hoarau et al., 2020). Telfer et al., 2010, investigated the interactions

between *Ap* and *Bartonella* sp, in natural populations of hosts finding that infection with *Ap* reduces the likelihood of *Bartonella* sp. infection by a factor of 0.35x. This negative association between the microparasites at first seems unusual, especially when considering both organisms occupy different ecological niches, the granulocytes and erythrocytes respectively. However, given the increased likelihood of coinfection of *Ap* with *B. microti* (up to ~5x) and the ability of *Ap* to reduce erythrocyte numbers, it is possible that bartonellae are unable to establish infections as effectively due to the depletion of their primary niche (Diuk-Wasser & EPJ, 2016). This work highlights the importance of viewing microbes in the context of their communities and demonstrates the complex interactions between microparasites of the blood.

One of the striking differences between *Anaplasma* and *Bartonella* lies in their cultivation generating significant challenges in a laboratory setting. *Anaplasma* species, specifically *Ap*, notorious for its obligate intracellular nature, poses formidable difficulties in conventional cultivation methods. *Ap* necessitates specialised cell culture techniques which restricts our ability to practically culture and sequence epidemiologically relevant strains (Salata et al., 2021). *Bartonella* sp. on the other hand can be cultured more readily, although not without challenges, namely their fastidious nature and specific nutritional requirements. These differences emphasise a need for innovative approaches to studying hard-to-culture bacteria to better understand their biology and facilitate advancements in disease prevention, diagnostics and therapeutics. Chapter 1 will introduce both the *Bartonella* genus and *Ap* discussing aspects of their biology, epidemiology and evolutionary relationships.

## 1.1: Classification & Characteristics of Bartonellae

The *Bartonella* genus is one of both academic and socioeconomic importance (Okaro et al., 2017). *Bartonella* are Gram-negative, slow growing, fastidious, facultative pathogens that are vectored by sandflies, lice, and keds and hosted by a diverse range of mammals (Roden et al., 2012). Humans are also common targets of various *Bartonella* species with notable zoonotic species being *Bartonella henselae* and *Bartonella quintana*, the agents of cat scratch disease and trench fever respectively (Anstead, 2016; Okaro et al., 2017; Mardosaite-Busaitiene et al., 2019).

In addition to these specialised zoonotic species, many other species of bartonellae have displayed a capacity for opportunistic infections in humans, with mild to moderate symptoms (Chomel et al., 2006). Rodents play an important role as reservoir hosts with more than 20 species being detected in blood samples across the globe (Okaro et al., 2017; Mardosaite-Busaitiene et al., 2019). The *Bartonella* genus underwent reform in the 1990's when the genera *Rochalimaea* and *Grahamella* were reclassified as members of *Bartonella* (Brenner et al., 1993; Birtles et al., 1995). Prior to this reclassification, the genus was exclusively populated by *B. bacilliformis*, and as of 2024 there are 39 extant species and 3 subspecies, although many more putative taxa remain partially characterised.

Phylogenetically, the *Bartonella* genus is split into four primary lineages (L1-L4). The bartonellae lineages are defined based on phylogenetic inference, genetic relatedness and the acquisition of specific virulence factors. Some lineages are associated with specific host tropisms and infection strategies. Typically, L1 and L2 species are limited to human and ruminant hosts respectively (Harms & Dehio, 2012; Minnick et al., 2014). Species of L3 and L4 can infect a much greater variety of vertebrate hosts, in part to the retention/ acquisition of the type four secretion system (T4SS) loci: *vbh*, *virB/D4* and *trw*. It is currently theorised that L3 and L4 had parallel adaptive radiations that were driven by the acquisition of the *virB/D4* T4SS and its primordial effector, leading to their uniquely diverse lineages and host specificities (Saenz et al., 2007; Engel et al., 2011).

The *Bartonella* genus is split into several lineages primarily based on phylogenetic, genomic and phenotypic characteristics. Whole genome comparisons provide the highest fidelity when interrogating the genus and therefore provide the best prediction for the evolutionary relationships between species and strains (Figure 1). In the absence of whole genome data, bartonellae can be characterised based on partial sequence data, for example a single locus such as *groEL* or through MLST if data is available (La Scola et al., 2003; Zeaiter et al., 2002). Unfortunately, the use of 16S rRNA to delineate the genus has proved to be unsuccessful due to high similarity across species (Kosoy et al., 2012). Phenotypic strategies are often useful for identifying isolates to the genus level, with electron microscopy providing a

detailed account of extracellular structures such as flagella and cell sizes which can vary across species (Gutiérrez et al., 2020).

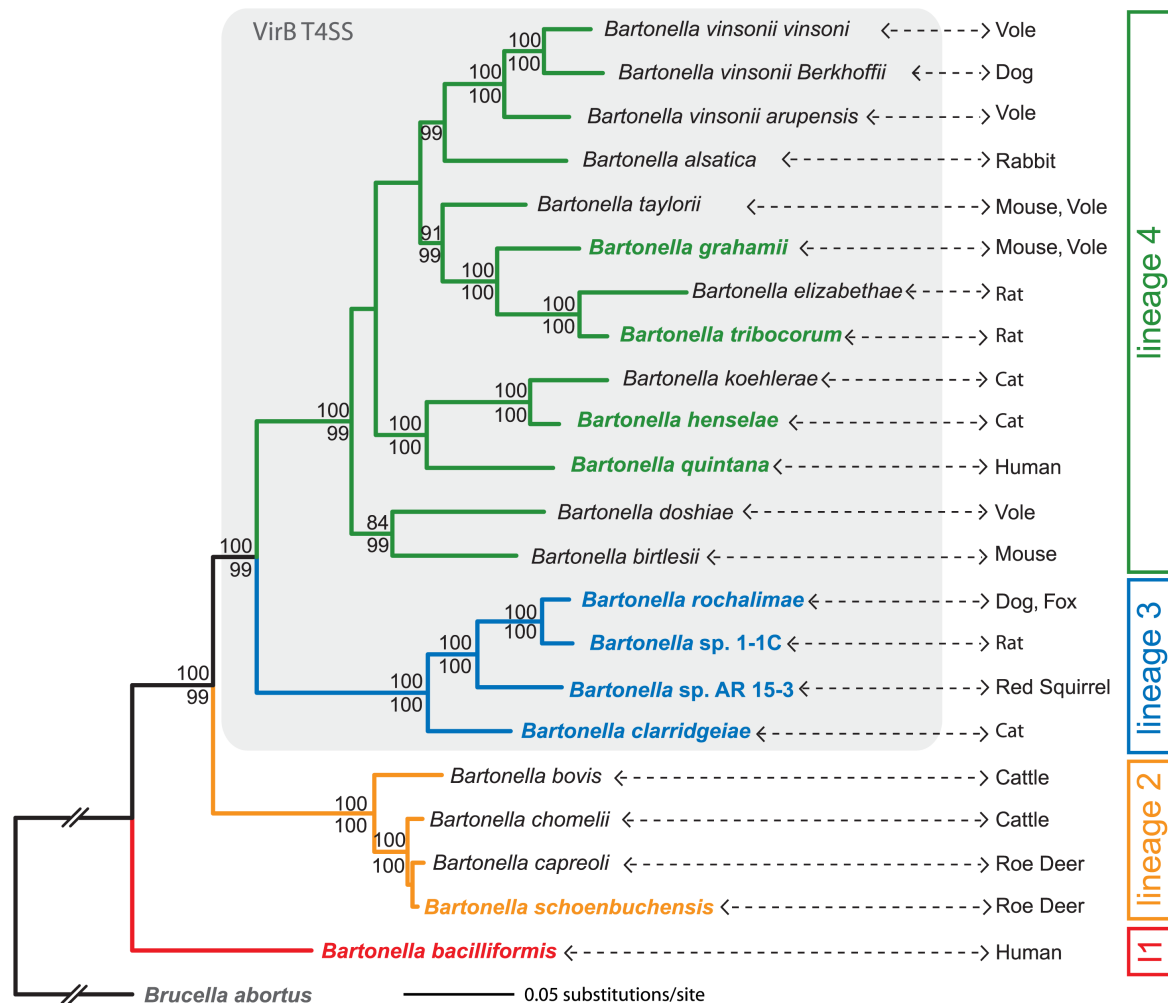
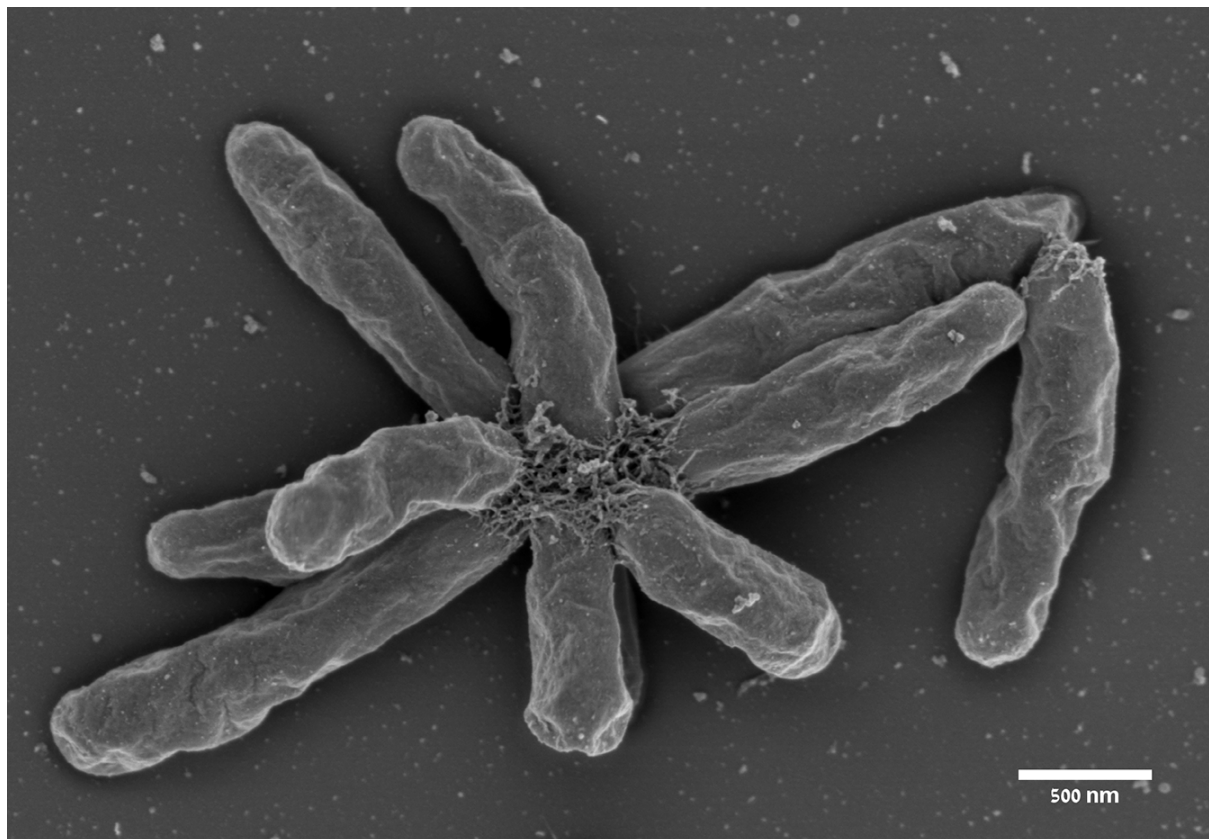


Figure 1: The bartonellae lineages delineated using core genome concatenations and maximum likelihood inference (Engel et al., 2011).

*Bartonella* species require haem as an essential nutrient as they lack the ability to produce protoporphyrin IX (PPIX) or haem (Fe<sup>2+</sup>-PPIX). Primary sources of haem include haemin (Fe<sup>3+</sup>-PPIX), haemoglobin, and therefore host erythrocytes. In contrast to many pathogens, *Bartonella* are also incapable of exploiting haem-rich scavenger molecules within the host, such as lactoferrin or transferrin (Minnick & Anderson, 2015). Whilst various liquid media have demonstrated a capacity for supporting *Bartonella* replication (Huang, 1967; Mason, 1970; Maggi et al., 2005), solid agars consistently offer the most dependable substrates for routine growth and contamination monitoring. Therefore, laboratories frequently employ chocolate agar, haemin agar or assorted blood agars to isolate and grow *Bartonella* species. Various intricate media have served as the foundation in these formulations, such as brain

heart infusion or Columbia agars. Supplements for these agar bases can differ but typically incorporate erythrocytes from accessible sources of defibrinated blood, including sheep, rabbit or horse. It is recommended that a neutral pH of 7 is maintained in formulation for optimal growth of the bacterium (Minnick & Anderson, 2015). The majority of *Bartonella* species exhibit both microaerophilic (require a minimal amount of oxygen for growth) and capnophilic (enhanced growth in the presence of elevated atmospheric CO<sub>2</sub>) characteristics that are usually established using a CO<sub>2</sub> incubator. Most species of *Bartonella* grow optimally at 35-37°C in a humidified, incubator with 5% CO<sub>2</sub>, however *B. bacilliformis* is a notable exception, requiring a cooler temperature of 25-30°C and ambient CO<sub>2</sub> for growth (Minnick & Anderson, 2015). This is curious as *B. bacilliformis* is a known human pathogen, which would suggest optimal growth at around 37°C. However, *B. bacilliformis* is widely considered the most ancient species within the genus retaining some ancestral traits that may favour this lower temperature. Additionally, *B. bacilliformis* has a far more limited geographical range, vectored by the sandfly, which has an environmental temperature range of 25-30°C suggesting increased adaptation to the vector rather than the human host (Sanchez et al., 2012). *Bartonella* colonies are initially very small in size (pinpoint), growing up to 1mm in diameter after several days or replication. Colonies have a round and lenticular shape, an even margin, and range from translucent to opaque. Colony colour can vary but typically embodies a creamy-white or light-brown shade depending on the age of the culture. Colonies may present as smooth or rough, and low passage isolates will grow slowest. Several passages can give rise to faster-growing bacteria that present as smooth which may indicate a competitive advantage in culture for the smooth morphotype (Kyme et al., 2003). Primary isolates may take weeks to become visible, but subsequent passages are known to grow much faster (within a week). *Bartonella* are Gram-negative, non-acid fast, and display a pleomorphic rod morphology. Staining can be achieved with Giemsa or Gimenez stains, while safranin is notably less successful. Typically, species of *Bartonella* present as coccobacilli or slightly curved rods with discernible polar enlargements. Cells may manifest as coccoid, beaded, filamentous or arrange themselves into chains, and some species of *Bartonella* are able to generate flagella and pili. Cells consistently measure less than 3µm, with a common size of 0.5µm in width by 1µm in length.



**Figure 2: Scanning electron microscopy of *Bartonella kosoyi* Tel Aviv strain, showing rod shaped bacteria interconnected by polar pili with a 500nm scale bar for reference (Gutierrez et al., 2020).**

*Bartonella* shares similarities in cell architecture and composition with other Gram-negative bacteria (Kreier et al., 1992). However, the genus exhibits several noteworthy and atypical features. One distinctive characteristic is the composition of *Bartonella* CAMP-like factor autotransporters (CFAs), where approximately half consist of cis-11-octadecanoic acid. Early studies also revealed a 'deep-rough' chemotype in the lipopolysaccharide (LPS) of *B. bacilliformis* and *B. quintana*, indicating minimal to no O-chain polysaccharide presence (Knobloch et al., 1988; Minnick, 1994; Liberto & Matera, 2000). A more recent biochemical analysis of *B. henselae* LPS uncovered a deep-rough molecule lacking an O-chain, featuring a unique inner oligosaccharide core comprising two KDOs (keto-deoxyoctulosonic acid) and one glucose molecule (Zahringer et al., 2004). The lipid A backbone in *Bartonella* is also atypical; instead of the typical glucosamine sugars, it contains a 2,3-diamino-2,3-dideoxy-glucose disaccharide, and both sugars are phosphorylated. Notably, the lipid A includes a long-chain fatty acid (C26:0 or C28:0), a feature observed in other intracellular pathogens like *Legionella* and *Chlamydia* (Zahringer et al., 2004). Studies indicate that *Bartonella* endotoxin is considerably less toxic



compared to LPS from other Gram-negative bacteria. Its activation of TLR-4 pathways is significantly lower (1000–10000-fold) than Salmonella LPS (Zahringer et al., 2004). Recent research even suggests that *B. quintana* LPS serves as a potent antagonist to TLR-4-mediated activation of human monocytes by *E. coli* LPS (Popa et al., 2007).

As *Bartonella* are microaerophilic, non-fermentative, and possess a relatively unremarkable physiology, conventional biochemical tests are less effective at identifying *Bartonella* at the species level. Nevertheless, peptidase tests assessing activity against L-proline and L-lysine have demonstrated utility in the presumptive and differential identification of specific pathogenic *Bartonella* species (Minnick & Anderson, 2015). These tests are often conducted using multi-test formats designed for anaerobes, such as the MicroScan Rapid Anaerobe identification panel. When it comes to novel species identification, many groups utilise non-specific API strip tests or assess other characteristics such as major fatty acid methyl ester compositions to delineate strains and species (Gutierrez et al., 2020). Biochemical tests are generally not ideal for identifying or delineating *Bartonella* and as a result, molecular techniques such as polymerase chain reaction (PCR) assays and DNA sequencing have become the preferred methodologies. PCR assays are a powerful tool for rapid identification of *Bartonella* species and a viable alternative to culture or serological approaches. Molecular detection of *Bartonella* has emerged as both a powerful diagnostic tool but also a tool for rapidly screening isolates to identify species of interest and novel variants. Common targets of assays include the 16S rRNA gene, the citrate synthase gene (*gltA*) (Birtles & Raoult, 1996) serine protease (*hrtA*) (Anderson et al., 1994), riboflavin synthase (*ribC*) (Bereswill et al., 1999), divisome protein (*ftsZ*) (Kelly et al., 1998), NADH dehydrogenase gamma subunit (*nuoG*) (Colborn et al., 2010), RNA polymerase beta subunit (*rpoB*) (Renesto et al., 2000), haem-binding protein A/Pap31 (*hbpA/pap31*) (Rolain et al., 2003), tmRNA (*sseA*) (Diaz et al., 2012), and the 16S-23S internal transcribed spacer (ITS) (Minnick & Barbian, 1997). Combinations of these genes have been utilised for multi-locus sequence typing (MLST) studies to unveil genetic variation across the genus and identify novel strains and species (Bai et al., 2013; Vayssier-Taussat et al., 2016).

## 1.2: Hosts, Reservoirs & Vectors

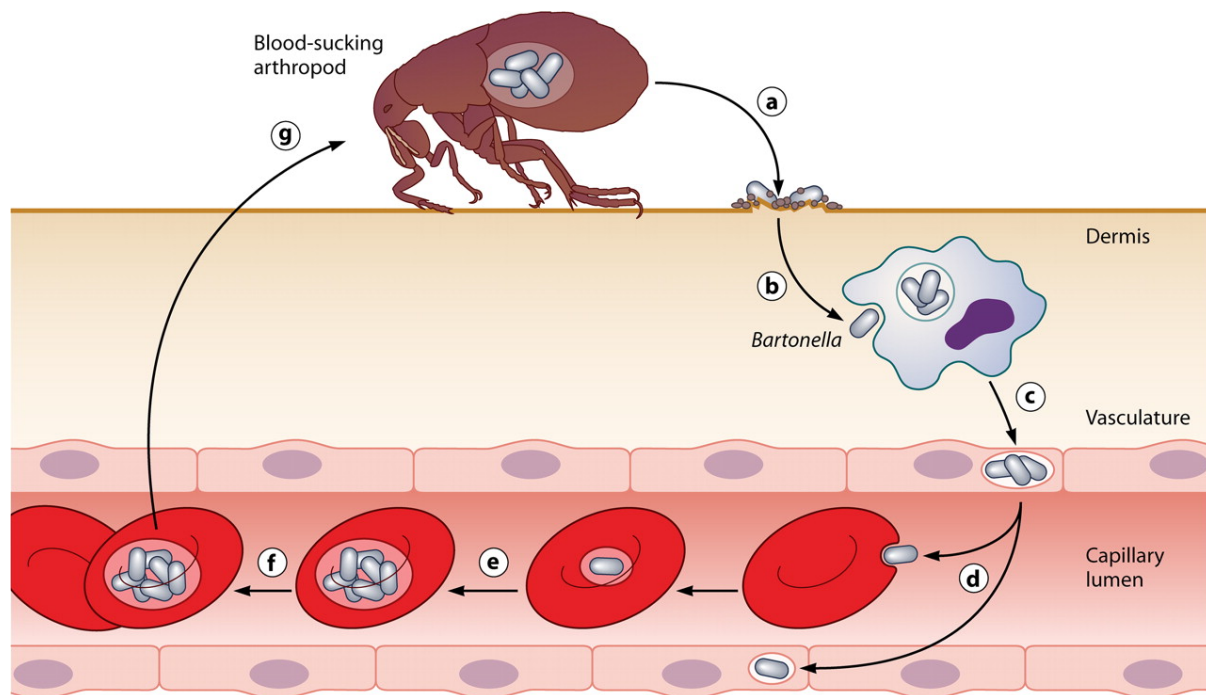
*Bartonella* are exceptionally successful bacterial parasites indicated by their ability to infect a wide variety of mammalian species. Members of the *Bartonella* genus exhibit host specificity and the ability to coexist within the same mammalian communities and individuals. To date *Bartonella* infections have been encountered in seven different orders of Mammalia (Birtles, 2005). For example, *B. bacilliformis* and *B. quintana* naturally only infect humans, whereas *B. henselae* infects members of the family felidae which are also targets of other *Bartonella* species such as *B. clarridgeiae* and *B. koehlerae*. European woodland rodents such as voles and wood mice are capable of maintaining several species of bartonellae (Holmberg et al., 2003; Birtles et al., 1994). The existence of over 20 validly published and proposed species of bartonellae in small rodents alone indicates variants are not directly competing for resources and are occupying unique ecological niches (Morick et al., 2011; Gutierrez et al., 2015; Tolkacz et al., 2018).

To date the ability of numerous arthropods to act as vectors for *Bartonella* transmission has been established. Fleas, lice, sand flies and keds are the most common culprits for *Bartonella* transmission, and as a result of the global distribution of these key vectors, species of bartonellae have become endemic across the globe. Notably, *Bartonella* exhibit a degree of vector specificity, so not all species can be detected in any single vector species. Rodents and their flea parasites have been demonstrated to carry a great diversity of bartonellae suggesting a high degree of specialisation among these organisms (Jones et al., 2008; Hawlena et al., 2013). Competence studies have identified both vertical and horizontal transmission pathways in flea communities and from fleas to rodents highlighting the importance of fleas as bartonellae vectors (Kosoy et al., 1998; Boulouis et al., 2001; Morick et al., 2013). Fleas are in fact so efficient at the harbouring and transmission of bartonellae that some suggest fleas represent an additional reservoir for suitably adapted variants (Birtles 2005; Deng et al., 2012). The transmission route from arthropod to mammal is achieved through the gastrointestinal content, mainly faeces (Birtles 2005). After ingestion of the blood meal, bartonellae will colonise the digestive tract of the arthropod vector. The bacteria will proliferate in the gut and digestive tract and are subsequently shed in the faeces, rather than being

transmitted through the salivary glands as seen in other vector-borne pathogens such as *Plasmodium* sp. and *Borrelia* sp. (Wells & Andrew, 2019; Shih et al., 2002). When the vector defecates near the feeding site in the fur, bartonellae are deposited into the environment. The host may then introduce the bacteria into its bloodstream or mucosal surfaces through scratching, grooming, or direct contact with the contaminated site. For example, the cat flea represents the primary vector of *B. henselae*. The competence of cat fleas in transmitting *B. henselae* has been experimentally established, playing a crucial role in sustaining infections within feline populations (Finkelstein et al., 2002). The faeces of fleas have been implicated as an important transmission route among hosts, as contaminated faeces in open wounds can establish infection (Bradbury et al., 2010). This method of transmission coined the name “cat scratch disease”, as a scratch from a cat infested with infected fleas can lead to human infection (Chomel & Kasten, 2010). The excreted flea faeces will contain *B. henselae* up to 24 hours after a blood meal and can establish bacteraemia in cats that can last for weeks (Bouhsira et al., 2013; Guptill, 2010). The human body louse represents the vector responsible for human-to-human transmission of *B. quintana*. Risk factors include unhygienic living conditions and lack of treatment for infestations. As a result, infections with *B. quintana* were severe in World War I and II trenches, and more currently in homeless populations (Regier et al., 2016). Sand flies transmit *B. bacilliformis* between humans. The prevalence of *B. bacilliformis* is exclusively restricted to the Peruvian Andes where the vector is widespread, but the impacts of climate change are a growing concern for extending the vectors habitat. Ticks are known vectors for many blood-borne pathogens such as *Anaplasma*, *Borrelia* and *Ehrlichia*. Various species of *Bartonella* have been detected in ticks across the globe where prevalence ranges from 0 – 80% depending on the region and species of tick investigated (Regier et al., 2016)

The natural cycles of bartonellae are surprisingly complex and can be divided into several stages as illustrated in Figure 2. *Bartonella* initiate their infection cycles within susceptible mammalian hosts after replication within the gut of arthropod vectors (Harms & Dehio, 2012). Upon feeding on mammalian blood, arthropods induce an inflammatory reaction and irritation in the host’s skin. This prompts the host to scratch the feeding site, and, in some cases, this may lead to the introduction of *Bartonella*-containing faeces into the dermis (Chomel et al., 2009). Following initial

inoculation into the dermis, the bacteria progressively enter dermal and blood-seeding niches. In the colonisation of the dermal niche, migrating cells such as dendritic cells or macrophages are likely co-opted by *Bartonella* to reach the blood seeding niche (Fromm and Dehio, 2021). At this stage the specialised T4SSs fulfil a crucial role in the migration of bartonellae from the dermis to the blood seeding niche as the repertoire of *Bartonella* effector proteins (beps) act as key virulence factors in the modulation of multiple host cell processes (Sorg et al., 2020). Once in the blood seeding niche, bacteria likely colonise the vascular endothelium, which then periodically releases bartonellae into the bloodstream (Eicher and Dehio, 2012). The vascular endothelium is certainly a strategic target for bartonellae due to three primary reasons: (1) its proximity to the bloodstream which will allow easy access to erythrocytes, the primary target of bacterial replication., (2) Endothelial cells provide a relatively immune-privileged niche, protecting bartonella from circulating immune defences (Resto-Ruiz et al., 2003)., (3) Endothelial cells have access to a large number of host derived compounds essential for survival and colonisation e.g. Iron and haem (Obino & Dumenil, 2019). Invasion of endothelium is thought to be achieved either by endocytosis of individual bacterium's or the formation of bacterial aggregates in the form of invasomes, however at this stage there is no conclusive evidence to support this theory (Truttmann et al., 2011). It is believed that bacteria released into the blood stream seek out and invade the erythrocytes using cellular systems such as flagella for motility or the trw T4SS for adhesion (Wagner & Dehio, 2019). Bacteria then replicate to an average of eight bacteria per erythrocyte after which apoptosis is induced (Pulliainen & Dehio, 2012). Bacteria released from the erythrocytes then seek out new erythrocytes for replication, explaining the occurrence of host bacteraemia shortly after infection (Harms & Dehio, 2012). The feeding of arthropod vectors on the infected mammalian host then transmits the bacterium back into the arthropod gut completing the cycle (Fromm & Dehio, 2021).



**Figure 3: The common strategy of Bartonellae to establish infection in reservoir hosts. (a) Bartonellae colonise the dermis (b) *Bartonella* co-opt migratory cells to the blood-seeding niche (c) Bartonellae establish infection in endothelial cells (d) Bartonellae are seeded into the blood stream where they invade the erythrocytes and other endothelial cells (e) Bartonellae replicate with host erythrocytes (f) inducing apoptosis (g) blood meals of the infected host transmit *Bartonella* back into arthropod vectors (Harms & Dehio, 2012).**

Throughout the course of infection, *Bartonella* benefits from the lack of an effective host immune response thanks to the globally moderated inflammatory profile induced by the highly effective immunomodulators released by the bacteria. It is worth noting however that different species of *Bartonella* possess slightly different strategies for host cell invasion and modulation due to the presence of distinct cellular mechanisms that have arose due to pathogen-host adaptation. For example, recent work by Fromm et al., 2024 has demonstrated that the type III secretion system (T3SS) effector yopJ, encoded in many bartonellae variants may function as a T4SS effector in species of bartonellae aiding in immune modulation and evasion. Nevertheless, it is generally accepted that all members of the genus follow the overall concept of this infection cycle.

### 1.3: Molecular Basis of Bartonellae Parasitism

A large amount of work has been carried out to understand the molecular basis of the bartonellae life cycle and parasitic strategy using both natural and experimental animal and vector models. As mentioned above the inoculation of an infected host is

primarily mediated by the introduction of infected vector faeces into cuts or scratches on the skin. This is followed by the invasion of the vascular endothelium, then the infection of erythrocytes before uptake back into arthropod vectors.

The extracellular matrix of endothelial cells acts as a barrier for bartonellae, which has led to several adaptations that facilitate both adhesion and degradation of the extracellular matrix. To efficiently adhere to the extracellular matrix *Bartonella* express surface adhesins including the well-studied adhesin A (BadA) and other trimeric autotransporter adhesins (Reiss et al., 2004). The degradation of the extracellular matrix is then facilitated by the interaction between fibrinolysis and pathogen proteins (Lahteenmaki et al., 2001). Alpha-enolase a metabolic enzyme produced by *Bartonella* to act as a plasminogen receptor mediates plasmin activation, which in turn promotes fibrin dissolution in the extracellular matrix of the endothelial cells (Keragala and Medcalf, 2021). This loosens or degrades the extracellular matrix of endothelial cells facilitating the entry of Bartonellae.

The initiation of intra-erythrocytic bacteraemia is a characteristic feature of Bartonellae infection in mammalian hosts. The process of infecting erythrocytes by Bartonellae can be split into three primary stages: adhesion, erythrocyte deformation, and invasion, all of which include diverse virulence factors that aid in host immune modulation and evasion (Jin et al., 2023). The exact mechanism for the invasion of mature erythrocytes (the primary target of bartonellae) is not fully understood, however, some of the essential factors have been identified. Adhesion of *Bartonella* to erythrocytes is achieved either via increased motility due to the expression of flagella, or the use of specialised trw T4SSs (Larrea et al., 2013). Remarkably, the acquisition of the trw T4SS in L4 species of *Bartonella* appears to have directly replaced flagella, with the expression of the systems mutually exclusive (Harms & Dehio, 2012). Upon attachment to the host erythrocytes our understanding wanes; invasion is thought to be achieved with the invasion-associated locus B genes (*lalB*), deformin, and haemolysin. *lalB* is a 19.9kDa protein with putative signal peptides, that have been demonstrated to confer erythrocyte invasion in experimental *E. coli* when knocked in and out of the genome (Mitchell & Minnick, 1995). Deformin on the other hand is known to mediate the production of trenches, pits, conical invaginations, and internal vacuoles in the erythrocyte membrane that aid in cell invasion (Benson et al., 1986). The identity and function of deformin in

*Bartonella* does however require further investigation with more recent studies suggesting multiple proteins in the supernatant may be involved (Hendrix & Kiss, 2003). Haemolysins have been identified in multiple species of *Bartonella* with notably two types, contact-dependent haemolysins and autotransporter cohaemolysins (Hendrix, 2000; Litwin & Johnson, 2005). Contact-dependent haemolysis is maximally expressed during the exponential growth phase of *B. bacilliformis* and may confer escape from the vacuoles or erythrocytes during intracellular parasitism, similarly, cohaemolysins are implicated in the escape of erythrocytes via cell lysis (Hendrix, 2000; Litwin & Johnson, 2005).

Erythrocytes notably lack the major histocompatibility complex (MHC) on the cell surface which prevents antigen presentation. This makes MHC-dependent cytotoxicity an ineffective response to *Bartonella* invasion. The host instead employs members of the innate immune system, macrophages and dendritic cells to clear the infection. Using *B. henselae* as a model, it was demonstrated that macrophage cell line J774 could rapidly internalise the bacterium. Within a span of 4 hours, unstimulated murine macrophages reached full saturation in phagocytising *B. henselae*, resulting in a notable elevation in the expression levels of tumour necrosis factor  $\alpha$  (TNF- $\alpha$ ), interleukin-1 $\beta$  (IL-1 $\beta$ ), and interleukin-6 (IL-6) by J774 (Musso et al., 2001). After phagocytosis of *B. henselae*, dendritic cells had elevated levels of the aforementioned cytokines and the chemokines CXCL8, CXCL1, and CXCL13 which play a critical role in recruiting neutrophils and B cells to the infection site (Vermi et al., 2006). The release of pro-inflammatory cytokines contributes to the formation of distinctive granulomas in cat scratch disease, facilitating the containment of *B. henselae* to specific sites within the host, thus preventing dissemination. Using mice as infection models it has also been demonstrated that antibodies play a significant role in the clearing of *Bartonella* cells in the blood (Koesling et al., 2001). Several studies have demonstrated that antibodies can prevent bacterial attachment to erythrocytes independent of complement or Fc receptors (Pulliainen et al., 2012). IgG and IgA antibodies are dominant during the course of cat scratch disease whereas IgM is dominant during the acute phase of *B. bacilliformis* infection (McGill et al., 1998; Pons et al., 2017).

*Bartonella* like other pathogenic bacteria are however in a constant arms race with their reservoir hosts, which has led to the adaptation of several fascinating strategies

to evade the immune response. Firstly, the evasion of the innate immune system is a prerequisite for successfully establishing infection. *B. tribocorum* is notably resistant to macrophages in rats, a mechanical resistance likely derived from BadA and other bacterial aggregates, which have been shown to reduce susceptibility to macrophages when compared to gene knock-out strains (Hong et al., 2017; Riess et al., 2004). Even when engulfed by phagocytes, pyroptosis can be suppressed, limiting the scope of an inflammatory response. Additionally, *Bartonella* can form unique *Bartonella*-containing vacuoles (BVC) that can also delay lysosomal targeting and destruction (Kyme et al., 2005). *Bartonella* have also adapted to possess a highly mutated lipopolysaccharide component on the outer membrane of their cells, that can no longer be recognised by Toll-like receptor 4 (TLR4), significantly contributing to a lower efficiency of phagocytosis (Popa et al., 2007; Mosepele et al., 2012; Malgorzata-Miller et al., 2016). This means that the host immune response is instead triggered by TLR2 recognition, thus avoiding the release of inflammatory cytokines triggered in the TLR4 pathway (Vermi et al., 2006; Matera et al., 2008). In *Candida albicans* TLR2 recognition has in-fact been shown to induce IL-10 secretion and promote Treg cell survival which instead inhibits an inflammatory response (Netea et al., 2004).

---

## 1.4: *Bartonella* Genome Evolution

The *Bartonella* genus now contains 39 validly published species and 3 subspecies with many more awaiting characterisations. Table 1 describes the general characteristics of the validly published *Bartonella* species that have publicly available whole genome data. Genomes range in size from 1.4 – 2.6Mbp with a gene content directly tied to size, with typically one gene per 1kbp. GC content remains relatively low throughout the genus ~37 – 39% with the exception of *B. apis*, which boasts a GC content of just over 45%. Some species of *Bartonella* have been observed to contain plasmids, which in some cases have been chromosomally integrated. *Bartonella* genomes contain several notable macromolecular systems including T4SSs, flagella, the *Bartonella* gene transfer agent (BaGTA), gene transfer agents such as *BadA*/ *Vomp*, and bacteriophages that may have important roles in virulence and ultimately the success of *Bartonella* infections (Guy et al., 2013).

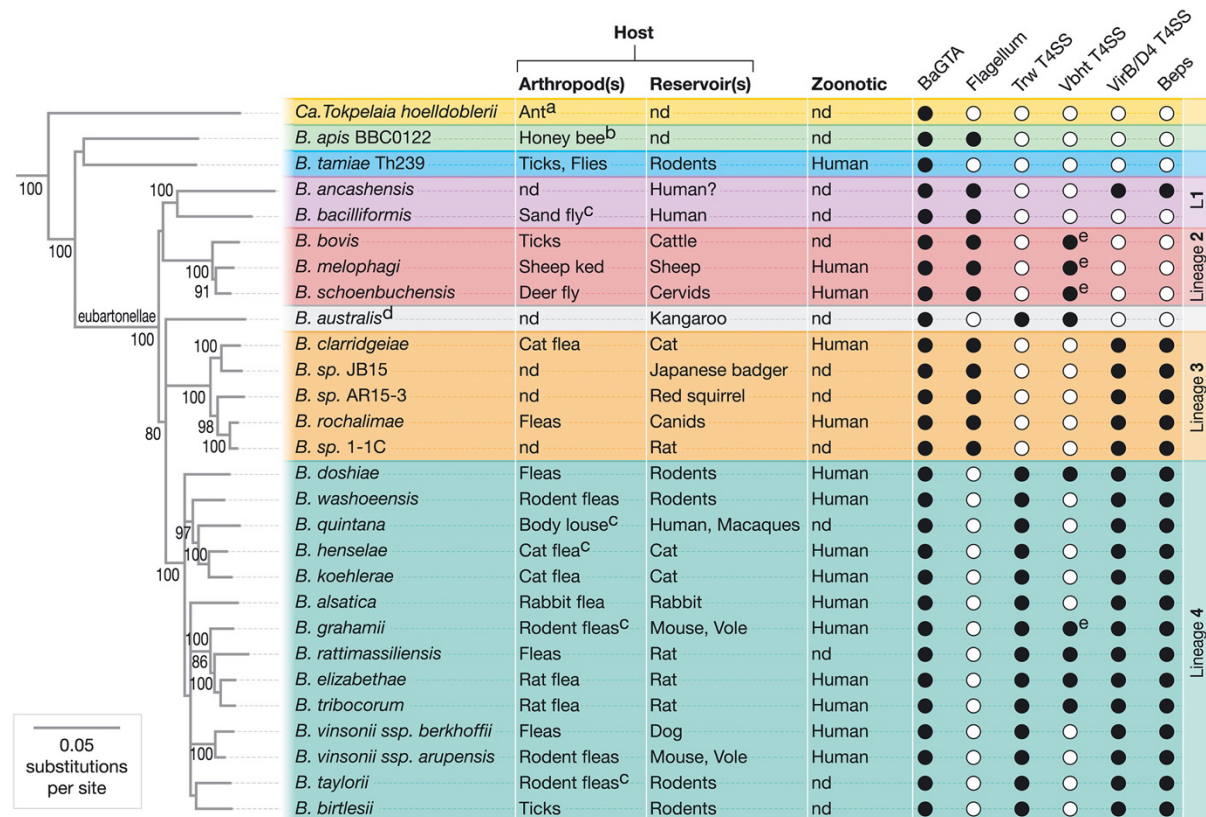


**Table 1: A collection of the best genome representations of all validly published species of *Bartonella* that have available data; retrieved from NCBI GenBank**

Accession	Species	Strain	Lineage	GC%	Size	Fragments	CDS	Plasmid Size
GCA_013388295.1	<i>Bartonella alsatica</i>	CIP 105477	4	36.85	1,659,117	1	1543	n/a
GCA_001281405.1	<i>Bartonella ancashensis</i>	20.00	1	38.42	1,467,695	1	1346	n/a
GCA_001952075.1	<i>Bartonella apis</i>	PEB 0122	n/a	45.45	2,603,122	16	2671	n/a
GCA_000015445.1	<i>Bartonella bacilliformis</i>	KC583	1	38.24	1,445,021	1	1365	n/a
GCA_000273375.1	<i>Bartonella birtlesii</i>	IBS 325	4	37.74	1,834,009	1	1887	n/a
GCA_000384965.1	<i>Bartonella bovis</i>	91-4	2	37.36	1,624,667	1	1473	n/a
GCA_014203215.1	<i>Bartonella callosciuri</i>	DSM 28538	4	38.42	1,745,037	36	1904	n/a
GCA_902813205.1	<i>Bartonella capreoli</i>	DSM 21569	2	37.83	1,814,765	439	2010	n/a
GCA_014138465.1	<i>Bartonella chomelii</i>	DSM 21431	2	37.53	1,565,755	33	1451	n/a
GCA_000253015.1	<i>Bartonella clarridgeiae</i>	73	3	35.73	1,522,743	1	1334	n/a
GCA_900445535.1	<i>Bartonella doshaiae</i>	NCTC 12862	4	38.00	1,844,489	2	1802	n/a
GCA_900638615.1	<i>Bartonella elizabethae</i>	NCTC 12898	4	38.37	2,018,538	1	2076	n/a
GCA_000312525.1	<i>Bartonella florencae</i>	R4	4	38.45	2,053,511	89	2217	n/a
GCA_014197255.1	<i>Bartonella fuyuanensis</i>	DSM 100694	4	36.67	1,944,744	81	1971	n/a
GCA_000022725.1	<i>Bartonella grahamii</i>	as4aup	4	38.04	2,341,328	1	2347	28,192
GCA_000046705.1	<i>Bartonella henselae</i>	Houston-1	4	38.23	1,931,047	1	1992	n/a
n/a	<i>Bartonella hiexiaziensis</i>	RE21	4	39.16	2,102,653	1	2124	29,248
GCA_000706625.1	<i>Bartonella koehlerae</i>	C-29	4	37.58	1,747,106	3	1739	n/a
GCA_003606325.3	<i>Bartonella kosoyi</i>	Tel Aviv	4	38.40	2,234,681	1	2332	41,483
GCA_003606345.3	<i>Bartonella krasnovii</i>	OE 1-1	4	38.10	2,157,564	1	2164	29,057
GCA_000312585.1	<i>Bartonella queenslandensis</i>	AUST/NH15	4	38.38	2,377,862	598	3138	n/a
GCA_000046685.1	<i>Bartonella quintana</i>	Toulouse	4	38.80	1,581,384	1	1652	n/a
GCA_000312565.2	<i>Bartonella rattaustraliani</i>	AUST/NH4	4	38.80	2,158,445	108	2628	11,227
GCA_902652675.1	<i>Bartonella refiksaydamii</i>	RSKK 19006	4	38.49	1,924,860	113	2044	n/a
GCA_000706645.1	<i>Bartonella rochalimae</i>	BMGH	3	35.69	1,534,143	3	1397	n/a
GCA_002022685.1	<i>Bartonella schoenbuchensis</i>	R1	2	37.85	1,672,726	1	1658	55,761
GCA_000312545.1	<i>Bartonella senegalensis</i>	OS02	4	38.61	2,002,317	41	2172	n/a
GCA_023920085.1	<i>Bartonella taylorii</i>	IBS 296	4	38.85	1,948,309	1	1877	n/a
GCA_000196435.1	<i>Bartonella tribocorum</i>	CIP 105476	4	38.82	2,619,061	1	2849	23,343
GCA_000278235.1	<i>Bartonella vinsonii</i> ssp. <i>arupensis</i>	OK-94-513	4	38.61	1,786,506	6	1695	n/a
GCA_000341385.1	<i>Bartonella vinsonii</i> ssp. <i>berkhoffii</i>	Winnie	4	38.83	1,802,699	1	1883	n/a
GCA_019659805.1	<i>Bartonella raoultii</i>	094	4	37.00	1,952,106	30	1714	n/a

Bacterial virulence factors aid in replication and dissemination of the bacterium through the subversion, evasion and modulation of host defences (Cross, 2008). T4SSs are a model example of bacterial virulence factors as these protein complexes are capable of transporting DNA, proteins, or effector molecules from the cytoplasm into the extracellular space (Wallden, Rivera-Calzada, & Waksman, 2010). T4SS are found in both gram-positive and gram-negative bacteria, across a

wide range of genera. The *Bartonella* genus has acquired three variations of T4SSs: vbh, virB/D4 and trw (Engel et al., 2011; Wagner & Dehio, 2019).



**Figure 4: Phylogenetic tree of the *Bartonella* genus based on conserved core genome homology. Vector, reservoir zoonotic capabilities and virulence factors present in each species is indicated (Wagner & Dehio, 2019).**

The virB/D4 locus is the most widely studied of the three and can be found in L3 and L4 *Bartonella*. It is theorised that two independent acquisition events in L3 and L4 ancestors resulted in the evolution of the virB/D4 T4SS and the *Bartonella* effector proteins (beps) we see in extant species (Saenz et al., 2007; Engel et al., 2011).

Beps are thought to have evolved from a single primordial effector that has undergone functional diversification and duplication. The result of this diversification can be seen in the vast repertoire of beps, the primary secretion of virB/D4 T4SS in *Bartonella* (Engel et al., 2011). Research shows that beps typically function through the manipulation and suppression of host cells, aiding in the evasion of the hosts immune system (Engel et al., 2011; Harms et al., 2017). It is thought that the immunosuppressive capabilities of L3 and L4 *Bartonella* sparked the adaptive radiations of these lineages thanks to greater host adaptability (Guy et al., 2013). Beps consist of an N-terminal effector domain, a central connecting oligonucleotide

binding fold (OB) and a C-terminal bipartite signal (Engel et al., 2011; Siamer & Dehio, 2015; Fromm & Dehio, 2021). The primary effector domain in Beps functions to facilitate post-translation modifications (PTMs) of target proteins, known as the FIC domain (Harms et al., 2016). The C-terminal BID domain (Bep intracellular delivery) comprised of a positively charged C-terminal stretch functions to aid in the translocation of molecules thorough interactions with type IV coupling protein (T4CP) (Schulein et al., 2005). This FIC-OB-BID architecture is conserved across beps suggest evolution from a single ancestral effector via independent gene duplication and recombination events. Diverse repertoires have arisen in three separate lineages (L1-3) and have become a crucial tool for bartonellae success (Fromm & Dehio, 2021).

The Vbh T4SS represents a homologous example of the virB/D4 T4SS and can be associated with the T4CP *TraG*. This secretion system can be either chromosomally integrated or exist on plasmids of which various haplotypes has been described. In L2 the vbh T4SS has been described as a virulence factor that much like virB/D4 has a functional molecule for secretion: VbhT (Harms & Dehio, 2012). Interestingly, the Vbh secretion system encoded on the plasmid pVbh in a strain of *B. schoenbuchensis* has been identified as a classical conjugation system rather than an immune modulator (Harms, Liesch, et al., 2017). Furthermore, the role of *VbhT* is suggested to act as an interbacterial effector due to being secreted secondarily to the relaxase-ssDNA substrate into recipient bacteria (Harms, Liesch, et al., 2017). *VbhT* has demonstrated an ability to covalently modify and deactivate the type II topoisomerases, gyrase and topoIV, although the consequences of these modification have not been quantified (Harms et al., 2015; Harms, Liesch, et al., 2017). Chromosomally integrated Vbh T4SS on the other hand have seen accumulations of deleterious mutations which have left the critical components *traG* and *traA* missing. Consequently, the chromosomally integrated counterparts are largely considered genetic remnants of a once functional system that is no longer useful (Harms, Liesch, et al., 2017).

The trw T4SS is restricted to L4 *Bartonella* where it has become a direct replacement for flagella, aiding in erythrocyte adhesion (Deng, Le Rhun, Le Naour, Bonnet, & Vayssier-Taussat, 2012; Vayssier-Taussat et al., 2010). Interestingly, the acquisition of this system from conjugative plasmid R388 has driven an increased

capacity for host adaptation, exemplified by the adaptive radiation within L4 (Harms & Dehio, 2012). Functionally, the acquisition of this system provides an effector protein delivery system (secretion of beps), a system that enables interactions with host cell membranes that can undergo co-evolution with hosts, whilst also providing adherence and invasion capabilities. Ultimately, it is thought that the functional specificity of beps as well as their interaction with host receptors and immune responses, drives the increased host ranges seen in L4. Unlike other Type IV secretion systems (T4SSs) within *Bartonella*, the trw system lacks a T4CP, a critical component for substrate translocation. Surprisingly, the species-specific infection of erythrocytes by the Trw-T4SS does not depend on effector translocation but rather relies on the extracellular exposure of variable pilin subunits, specifically *TrwL* and *TrwJ* (Deng et al., 2012; Vayssier-Taussat et al., 2010). Notably, a *TrwJ* paralogue from the mouse-specific pathogen *Bartonella birtlesii* exhibits the ability to bind to mouse erythrocytes but not to cat erythrocytes. Studies have demonstrated that a *TrwJ* paralogue binds to the major glycoprotein band3 on the surface of erythrocytes (Deng et al., 2012). The Trw-T4SSs encode multiple variant copies of pilin subunits, resulting from gene duplication and diversification events. It is plausible that the polymorphic surfaces of erythrocytes have driven the diversification of these pilin subunits (Harms & Dehio, 2012). The Trw-T4SS serves as a crucial virulence factor facilitating reservoir-host-specific erythrocyte infection by L4 *Bartonella*, contributing to both adaptability and virulence in new mammalian hosts.

## 1.5: Classification & Identification of *Anaplasma phagocytophilum*

*Ap* is an obligatory intracellular Gram-negative bacterium of the family *Anaplasmataceae* in the order Rickettsiales. *Ap* has been described as the agent of tick-borne fever (TBF) in ruminants and granulocytic anaplasmosis (GA) in canines, equines and humans (Dumler et al., 2001; Dumler et al., 2005). TBF was first described in domestic livestock in the UK and mainland Europe where it is the most widespread tick-borne disease of animals (Stuenkel, 2007), and later described as GA in North America, where it is a commonly reported tick-borne infection of humans (Nathavitharana, & Mitty, 2015). *Ap* is now considered an emerging zoonotic parasite that has been detected in 33 countries, across 4 continents, primarily across the

northern hemisphere (Karshima et al., 2022). The classification of *Ap* has changed several times since it was first discovered in 1932 as an unknown tick-borne agent of fever in Scottish sheep (Gordon et al., 1932). The bacterium was formally characterised in 1951 as *Rickettsia phagocytophila*, then *Cytoecetes phagocytophila* in 1962, and *Ehrlichia phagocytophila* in 1974 (Dugat et al., 2015). Finally, two members of the genus *Ehrlichia*, *E. phagocytophila* and *E. equi*, the agents of human and equine granulocytic ehrlichiosis respectively were united under one name, *Anaplasma phagocytophilum* (Gribble, 1969; Chen et al., 1994; Dumler et al., 2001). Early attempts at studying the biology of *Ap* were significantly held back due to a lack of *in vitro* cultivation techniques. It wasn't until the mid 1990's that *in vitro* culture of *Ap* was achieved in HL60 cells (Goodman et al., 1996). Later researchers would develop Ixodid tick cell lines, namely, IDE8 and ISE6 derived from *Ixodes scapularis* that were capable of supporting commonly studied variants of the pathogen (Munderloh et al., 1996; Woldehiwet et al., 2002).

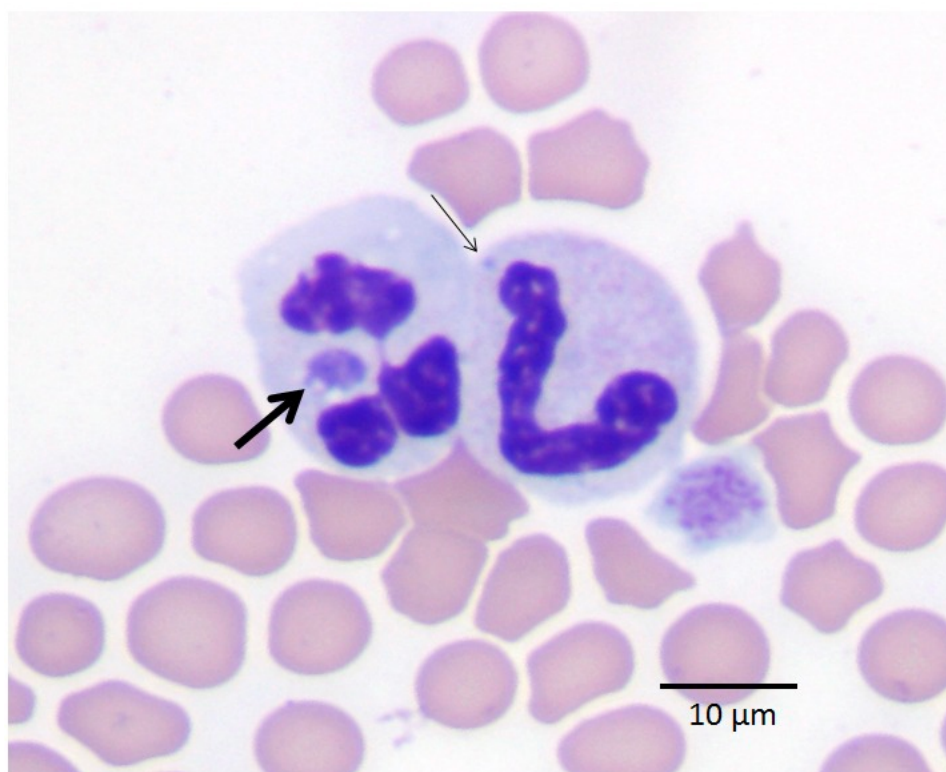


Figure 5: A peripheral blood smear from a cat infected with *Anaplasma phagocytophilum* stained using Wright-Giemsa stain. The larger arrow points to a morula within the cytoplasm of a neutrophil. The smaller arrow points to a Dohle body for comparison (Headquarters, 2016).

The first continuous Ixodid tick cell lines were established in 1975 from developing adult *Rhipicephalus appendicalatus* tissues (Varma et al., 1975) and are still in use

today (Bell-Sakyi et al., 2018). These first cell lines were shortly followed up by multiple cell lines that were instead derived from embryos of several *Dermacentor*, *Rhipicephalus* and *Hyalomma* species (Holman & Onald, 1980; Holman, 1981; Yunker, 1981; Yunker, 1984; Kurtti & Munderloh, 1982; Kurtti et al., 1982; Kurtti et al., 1983; Bell-Sakyi, 1991), most of which are currently extant (Bell-Sakyi et al., 2018). The most widely used tick cells line today were developed in the 1990's (Munderloh et al., 1994). These cell lines were derived from *Ixodes scapularis* and coincided with a renewed interest in tick-borne disease in North America (Spach et al., 1993). The tick cell lines IDE8 (Munderloh et al., 1994) and ISE6 (Kurtti et al., 1996) were subsequently distributed to research groups and extensively used to study tick-borne diseases due to their broad susceptibility to many species and strains (Oliver et al., 2015; Cabezas-Cruz et al., 2017). These two tick cell lines have played a pivotal role in researching *Ap*, being used to elucidate host-vector-pathogen relationships and provide a platform for generating whole genome data (Al-Rofaai & Bell-Sakyi, 2020). The molecular revolution was underway, and novel studies were beginning to describe elements of the *Ap* genome which spawned hypotheses on phylogeny, ecotypes and host specificity.

Because of the difficulties associated with isolation of *Ap*, it's detection and characterisation have, for the most part, been reliant on PCR-based approaches. Conventional PCRs (cPCR), nested PCRs (nPCR), and real-time PCRs (rtPCR) have all been extensively used by researchers to assess the infection intensity, incident rate and genetic variability of *Ap* populations (Karshima et al., 2022). Meta-analyses of these datasets have in fact revealed the epidemiological risk of the pathogen is notably underestimated given human encroachment into natural tick habitats, and the endemic nature of the pathogen (Pilloux et al., 2019). Early efforts at classifying *Ap* genetic diversity were aimed at generating 16S rRNA data from wild and domesticated animals (Hulinska et al., 2004). However, due to the high intra-species sequence similarity, there was insufficient data to identify the diversity of *Ap* in Europe. Investigating other molecular targets such as *msh2*, *groEL* and *ankA*, which would provide a notably better resolution for examining *Ap* diversity and epidemiology (Drazenovich et al., 2006; Jahfari et al., 2014; von Loewenich et al., 2003). These single gene studies would ultimately conclude that one gene is insufficient to determine the genetic diversity of a strain of *Ap*. However, several



insights could be gained, for example the *groEL* gene still provides some of the core principles of *Ap* diversity identifying the existence of four genetically distinct ecotypes that may have unique epidemiological cycles.

## 1.6: Hosts, Reservoirs and Vectors

Although *AP* has been detected in several hard tick genera including *Amblyomma*, *Dermacentor*, *Haemaphysalis*, *Hyalomma*, *Ixodes* and *Rhipicephalus* (Karshima et al., 2022), *Ixodes* sp. ticks are undoubtedly its most important vector (Telford et al., 1996; Ogden et al., 2002). The genus *Ixodes* is associated with a variety of bacterial, viral, and protozoan pathogens of veterinary and medical importance across the globe including the agents of Lyme borreliosis, babesiosis, and tick-borne encephalitis (Karshima et al., 2022). *Ixodes* ticks can be found in almost all geographical regions of the world due to the impressive survivability of the different species (Xu et al., 2003). The geographical distribution of *Ixodes* ticks can be correlated with the distribution of *Ap*, with a global meta-analysis of *Ixodes ricinus* complex ticks indicating an infection rate of between 6.84 – 9.31%, depending on the species in question (Karshima et al., 2022). This makes the distribution of *Ap* particularly susceptible to the impacts of climate change, changes in land use, urbanisation, international trade of wild and domestic animals, and changes in ecosystems (Daszak et al., 2001). In particular, longer wetter summers and shorter warmer winters are extending seasonal activity and territory of pertinent tick vectors, thus increasing the risk of infection worldwide (Gray et al., 2009; Gasmi et al., 2018; Sonenshine, 2018). As a result of the global distribution of *AP*, no single species of *Ixodes* can be singled out as the primary vector species. For instance, in Europe, *I. ricinus* is the primary vector, whereas in Asia, it is *I. persulcatus*, and in North America, *I. scapularis* (Bown et al., 2009; Stuen et al., 2013). Other notable vectors include *I. trianguliceps* in Europe and *I. pacificus* in the USA (Bown et al., 2008; Stuen et al., 2013).

The European sheep tick *I. ricinus* is characterised by a three-host life cycle and seasonality (Reye, Hubschen, Sausy & Muller, 2010). Development from larvae to mature tick is stimulated through feeding on trans migrant vertebrates in the open environment (Rymaszewska & Grenda, 2008). Due to the seasonal nature of the

ticks, development cycles usually span three years, however, completion of the cycle in two years or less is not unprecedented.

*Ixodes* ticks become infected with *Ap* through acquisition feeding on an infected vertebrate (Villar et al., 2016). This means an understanding of *Ap* host specificity is key to understanding at what stage of the tick's lifecycle pertinent strains of *Ap* are being introduced (e.g., via rodents in the early stages of development, or through larger herbivores once maturation is complete). There are several known factors that determine successful infection: the most important being the percentage of infected neutrophils within the vertebrate host, and the density of ticks feeding on this same host (Ogden, Casey, Woldehiwet & French 2003). Both transstadial and transovarial transmission have been implicated for maintaining *AP* within its endemic cycles, however, so far there is no evidence to suggest that transovarial transmission is possible within *I. ricinus* (Medlock et al., 2013; Jahfari et al., 2014; Krucken et al., 2013). The transmission and replication mechanisms of *AP* within tick vectors is poorly understood but studies have shown that *Ap* resides in the midgut and salivary glands of an infected tick (Sukumaran et al., 2006; Reichard et al., 2009). Within infected ticks, colonies of *A. phagocytophilum* have been predominantly observed within the hemocoel side of the midgut muscle cells and are thought to migrate to the salivary gland, muscle, and ganglia cells at a reduced frequency (Reichard et al., 2009). Vertical transmission is a distinct possibility for *Ap* within the tick vector. Studies have shown however that although possible vertical transmission of the bacterium was generally inefficient but could potentially be of use in environments where host interactions are limited (Krawczyk et al., 2022). Although rare, vertical transmission has been documented in mammals (i.e. from infected mother to offspring), but this is not considered a major route of transmission, with tick bites encompassing the primary transmission strategy (Krawczyk et al., 2022).

Imaging of *Ap* using transmission electron microscopes (TEMs) revealed two distinct morphotypes: reticulate and dense core (Reichard et al., 2009). The reticulate morphotype appears to act as a purely replicative, non-infectious form of *Ap*, whereas the dense-core cells retain the ability to infect nearby neutrophils whilst also exhibiting increased environmental resistance (Troese et al., 2011). The life cycle of *Ap* begins with the feeding of an infected tick vector on a susceptible vertebrate host.



*Ap* will be passed from the tick's salivary glands into the vertebrate's bloodstream (Matei et al., 2019). *Ap* then parasitises the neutrophils, and in rare cases the eosinophils through the formation of an intracytoplasmic inclusion derived from the cell membrane of the granulocyte by a dense-core cell (Rikihisa, 2010; Euroimmun, 2021; Langford Vets, 2021; Khatat, 2017). *Ap* then releases an arsenal of host manipulating molecules to establish infection (Alberdi, Espinosa, Cabezas-Cruz, De la Fuente, 2016). Amongst these strategies is the inhibition of the neutrophil apoptosis program through the manipulation of proapoptotic and antiapoptotic gene expression (Carlyon & Fikrig, 2006). This prolongs the life of an otherwise short-lived host cell, promoting increased replication and subsequent dissemination into the hosts bloodstream (Carlyon & Fikrig, 2006; Alberdi et al., 2016). Both reticulate and dense core morphotypes are thought to be capable of replication by binary fission to produce between 1 to 20 *Ap* cells each, however, some theories suggest only vegetative reticulate cells replicate to form morulae (Carrade, Foley, Borjesson & Sykes, 2009). Morulae are the result of excessive replication within host granulocytes; they are characterised as basophilic intracellular inclusions that mature into dense-core cells which are released through exocytosis or lysis of the granulocyte (Diniz & Breitschwerdt, 2012; Pruneau et al., 2014). Dense-core cells will then seek out passing granulocytes to replicate further.

*Ap* infections in livestock represents a huge veterinary burden for farmers inflicting significant economic losses whilst also impacting animal welfare (Lihou et al., 2020). In both sheep (*Ovis aries*) and cattle (*Bos taurus*), infection can lead to reduced yields, fertility and abortion storms in naïve animals (Lihou et al., 2020; Stuen et al., 2013). Infection with *Ap* can produce a variety of symptoms ranging from subclinical to fatal, making diagnosis particularly difficult without laboratory equipment (Bauer et al., 2021). Tick-borne fever (TBF) typically affects younger animals such as lambs, where the characteristic signs of infection are high fever, anorexia, dullness, nasal discharge, and lacrimal secretion (Grøva et al., 2011; Almazán et al., 2020). TBF alone is not a particularly deadly disease but does induce neutropenia and thrombocytopenia which can facilitate the acquisition of more aggressive secondary infections (Stuen et al., 2010; Gokce, & Woldehiwet, 1999). *Mannheimia haemolytica* and *Bibersteinia trehalosi* are common secondary infections in lambs and can result in significant respiratory distress (Daniel et al., 2017; Øverås et al., 1993). Moreover,

co-infection with staphylococcal bacteria in lambs may also result in tick pyaemia (TP) and severe polyarthritis (Sargison, & Edwards, 2009). A study conducted in the UK attempted to assess the impacts of TBF and TP finding that nearly half of the >30 million sheep in the UK lived in hilly, and often tick infested pasture, and estimated that more than 300,000 British lambs developed TBF followed by TP annually (Brodie, Holmes & Urquhart, 1986). Most lambs that develop TP are of no economic value and suffer throughout an often-shortened lifespan (Woldehiwet, 2006).

Roe deer and many other species of deer have been implicated as reservoir hosts for several strains of *Ap* in the UK and Europe (Robinson et al., 2009; Johnson et al., 2021; Woldehiwet et al., 2006). In the Netherlands, principal component analysis indicated that the density of *Ap* positive nymphs increased in forests occupied by red, fallow and to some extent roe deer populations (Takumi et al., 2021). The mechanisms behind this observation are still elusive but point to wild deer populations as reservoir hosts for *Ap*. Wildlife management rangers in Scotland have in fact found success when managing local deer populations either through fencing off woodland areas, or culling populations (Gilbert et al., 2012). They ultimately showed that effective wildlife management can reduce the impact of endemic tick-borne diseases through a reduction in the spread of vectors across wild habitats (Stuen et al., 2013). As such many farmers today and indeed for centuries have used preventative measurements such as avoiding tick infested pasture, spraying acaricides and protecting young animals until their innate immunity has fully developed to mitigate infection risk.

*Ap* infections in small mammals have been described both in the US and Europe (Foley et al., 2008; Bown et al., 2006). In Europe small mammal infections are thought to be delivered by a different Ixodid host, *I. trianguliceps* a small nest building hard tick (Bown et al., 2008). This set of epidemiologically separate strains have been characterised as ecotype III in Europe but have been encountered substantially less than ecotypes I and II (Jahfari et al., 2014). Bown went on to highlight common shrews (*Sorex araneus*) as one of the potential reservoir hosts for small mammal associated strains of *Ap* (Bown et al., 2011). Shrews were found to have a significantly higher chance of infection with 18.7% being infected, compared to 6.4% for field voles (*Microtus agrestis*) across more than 2000 samples over 2 years. Little is known about the genetics of UK *Ap* strains due to a lack of genomic

characterisation and less is known about the rodent associated strains of *Ap* that appear genetically distinct from other ecotypes, posing risks to disease management of potentially zoonotic strains. US rodent strains appear to be much more closely related to the predominant *Ap*-HA strains in the region and as such are treated as a potential threat to human health. Birds are associated with ecotype IV in Europe and are vectored by *Ixodes frontalis*, the bird tick (Jahfari et al., 2014).

Between 2001 and 2015 human granulocytic anaplasmosis (HGA) was reported over 15,000 times in the USA and in 2021 alone over 6700 times, a high number of cases when compared to Europe which reported fewer than 300 (Bakken & Dumler, 2015; Dahlgren et al., 2011; Matei et al., 2019). The exact reason for this disparity is unknown, but it is speculated that the primary cause is the circulation of a different subset of strains with more zoonotic potential in North America. North American strains are more virulent in humans producing more recognisable symptoms than their European counterparts with higher morbidity and mortality rates (< 1%); to date no fatal cases of HGA have been reported in Europe (ECDC 2022). It is feasible that there is greater HGA awareness amongst medical professionals in the USA, contributing to a larger number of reported cases. However, despite the few cases of HGA reported in Europe, human seroprevalence is relatively high. Antibodies to *AP* have been detected in 2-28% of the examined populations in various European countries, suggesting infections may not be as rare as they currently seem (De Keukeleire et al., 2017). This is especially curious when you consider that global seroprevalence of *Ap* ranged from 0-37% implicating some areas of Europe as relative hot spots for potential zoonotic infections (Wang et al., 2020). It is however important to note that HGA in Europe is especially known for high rates of asymptomatic infections, or cross-reactivity with other bacteria, pointing to a trend of incomplete diagnosis (Matei et al., 2019).

In most cases, clinical manifestations of HGA arise as a mild febrile illness, but can present with severe fever, chills, headaches, and myalgia. The most frequent abnormalities were thrombocytopenia, leukopenia, and increased levels of hepatic enzymes in serum (Rar, Tkachev, & Tikunova, 2021). In rare cases, HGA has been associated with meningitis, encephalitis and cerebral infarction (Bakken, & Dumler, 2015; Dahlgren et al., 2011; Kim et al., 2018). Interestingly, the acquisition of *Ap* is

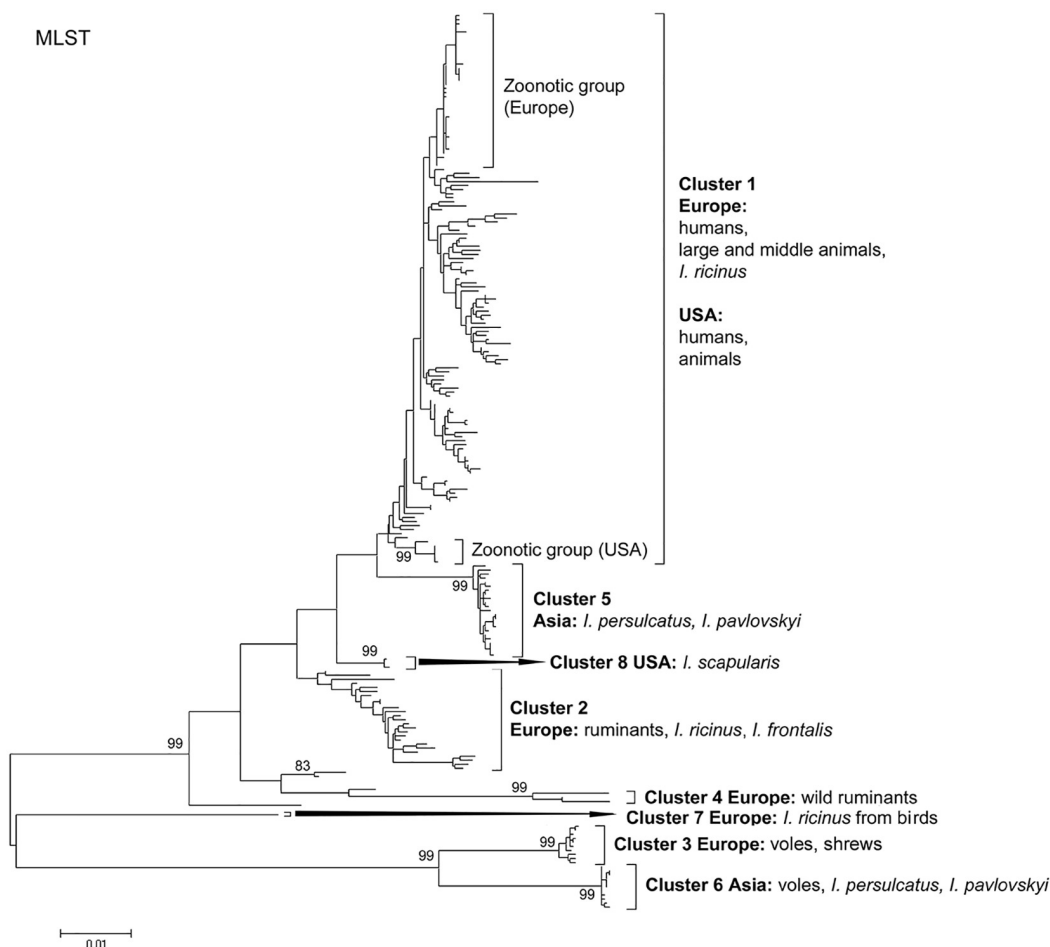
not limited to the bite of an infected tick but can be acquired both perinatally and via blood transfusions (Dhand et al., 2007; Annen et al., 2012).

## 1.7: Diversity & Strategies of *Anaplasma phagocytophilum*

The genetic diversity of *Ap* has been studied with a range of different molecular techniques. Initially, the highly conserved 16S rRNA gene was selected for multiple studies across both North America and Europe (Dumler et al., 2001; Massung et al., 2002; Rar and Golovljova, 2011). The study of the 16S rRNA gene in the USA revealed the circulation of two distinct groups that could be distinguished based on two substitution mutations. These are known as *Ap*-ha and *Ap*-variant 1, the human active variant and the non-human active variant respectively (Massung et al., 2002). European studies on Norwegian sheep revealed entirely novel variants both having different 16S rRNA sequences and pathogenic properties to their North American counterparts and each other (Stuenkel et al., 2002; Stuenkel et al., 2003). This made the *Ap*-ha and *Ap*-variant 1 delineations controversial, as the discriminatory power of 16S rRNA is too low for typing *Ap* (Langenwalder et al., 2020; Scharf et al., 2011; von Loewenich et al., 2003). Despite this, *Ap*-ha and *Ap*-variant 1 remained an important delineation of North American strain diversity due to its ability to predict pathogenic ability in humans. These findings promptly led to the search for both better single-locus and multi-locus targets. The most commonly used single-locus targets include the *ankA*, *groEL*, *gltA*, *msh2* and *msh4* genes which have had variable success in delineating *Ap* strains (Battilani et al., 2017; de la Fuente et al., 2005; Jahfari et al., 2014; Rar et al., 2014; Scharf et al., 2011; von Loewenich et al., 2003; Zhan et al., 2010). The *groEL* gene has been utilised for the delineation of *Ap* strains across the globe and particularly in Europe.

Multi-locus sequencing analysis includes techniques such as variable number tandem repeat (VNTR) and multi-locus sequence typing (MLST). Bown et al., 2007 first described a method based on microsatellite VNTRs, which were eventually proven to be too variable for phylogenetic analysis. However, a second technique was then developed by another group several years later which targeted minisatellite VNTR sequences, which were revealed to have less variation than Bown et al.,

2007's microsatellite targets, enabling the delineation of strains derived from different hosts possible (red deer and domestic ruminants vs. roe deer) (Dugat et al., 2014). Despite this success, VNTR studies were still less successful than comparable MLST studies. MLST studies targeted multiple loci such as *ankA*, *groESL*, *gyrA*, *msp4*, *pled*, *polA* *recG*, *typA* concatenating sequences in order to accumulate more genetic mutations for phylogenetic analysis, a now golden standard for phylogenetic studies of *AP* (Chastagner et al., 2014; Dugat et al., 2015). An MLST scheme developed by Huhn et al., 2014 which targeted seven housekeeping genes has played a major role in the study of *Ap* epidemiology, identifying eight distinct clusters segregated based on hosts, vectors, and geography (Figure 6). A more comprehensive compilation of MLST results for this typing scheme has been compiled at (<https://pubmlst.org/organisms/anaplasma-phagocytophilum>) totalling 724 isolates and 863 alleles. In total 9 clusters could be delineated from these isolates five of which have origins in Europe, two in the USA, and two in Asia. Cluster one accounts from the vast majority of isolates across both Europe and the USA including the primary zoonotic groups from both continents. Interestingly, the zoonotic groups from Europe and the USA appear to be genetically distinct using this typing scheme (despite belonging to the same cluster) implicating the human active US strains independently evolved. Curiously cluster 2 represents ruminant infections in Europe but also isolates of the bird tick *I. frontalis*. Cluster seven encompasses isolates derived from birds and *I. ricinus*. Clusters 3 and 6 suggest that European and Asian rodents are harbouring genetically related variants of *Ap* specifically adapted to infect small mammals. The results of this study imply *Ap* is a highly diverse pathogen with distinct host and vector preferences that can be traced geographically.



**Figure 6: Phylogenetic tree based on seven concatenated housekeeping gene sequences (2877 bp) of *A. phagocytophilum*. Phylogenetic analysis was performed with the maximum likelihood method based on the Tamura-Nei model in MEGA 6.0 with 1000 bootstrap replicates (Huhn et al., 2014).**

*Ap* is maintained through enzootic cycles between ticks and susceptible wildlife (Jahfari et al., 2014). *Ap* is currently believed to be a single species that is capable of infecting a variety of vertebrates and ticks (Chastagner et al., 2017). Despite this, *Ap* exhibits a capacity for host specialisation, clear when isolates from distinct host origins were not uniformly infectious for heterologous hosts (Jahfari et al., 2014). In addition to this, researchers were able to delineate *Ap* in clusters and subclusters based on molecular ecotyping (Jahfari et al., 2014). Ecotypes are defined as a population of cells in the same ecological niche which led to a delineation of strains based on molecular data and host tropisms. The *groEL* gene provides intermediate genetic variability and critically is able to delineate ecotypes better than other molecular targets such as 16S rRNA (Jahfari et al., 2014). A study conducted on 548 samples across Europe by Jahfari et al., 2014 popularised the idea of *Ap* ecotypes and set the foundation for following studies on *Ap* epidemiology. They found that four distinct and potentially epidemiologically separate ecotypes of *Ap* existed that have

distinct host tropisms. Ecotype I was the most common and encompassed the vast majority of livestock infections. Ecotype II was the second most common and is associated with primarily roe deer. Ecotype III was comparatively rare and had only been isolated from infected rodents and ticks. The final ecotype was only encountered a few times in birds and appeared genetically distinct from the other ecotypes. In total 97 haplotypes were observed indicating that *Ap* has significant global diversity. Ecotype I had the broadest host range but did notably lack rodents or birds, nonetheless, indicating it as a host generalist. This coupled with the generalist feeding nature of *I. ricinus* nymphs and adults likely facilitates the continuous exchange of ecotype I strains in the environment. All human cases can also be traced to ecotype I, suggesting the risk for zoonotic infections is high. We don't however see this in Europe, with the number of human cases being less than a few hundred (Matei et al., 2019). This opens up the possibility for further subdivisions within ecotype I, given the little amount of human data we have and the sheer diversity within the ecotype. More recently Jaarsma et al., 2019 and Grassi et al., 2021 built upon this, identifying subdivisions and novel haplotypes within the four ecotypes. Interestingly, the *groEL* ecotypes largely agree with Huhn et al., 2014's MLST scheme however there are some notable exceptions. For example, the MLST scheme identified an additional European ecotype comprised of wild ruminants; in addition to this, *I. frontalis* isolates were included in cluster 2 akin to ecotype II, which epidemiologically is difficult to justify. The ecotypes derived from *groEL* and Huhn's MLST scheme are still extensively used by researchers in Europe as there is little whole genome data to work with when constructing more robust representations of *Ap* evolution. As these two methodologies have fundamental delineations that disagree, it is essential that more whole genome data is generated in order to comprehensively explore strain diversity, host tropisms and the currently accepted ecotypes. The status of *Ap* populations in the UK has been extensively studied using the *groEL* gene, identifying the presence of three of the four European ecotypes in a vast range of vertebrate species; yet no complete representations of *Ap* genomes have been generated, limiting the scope of UK based epidemiological studies (Aparicio et al., 2023; Bianchessi et al., 2023; Gandy et al., 2022).

To date there are 33 complete genomes of *AP* available across both the NCBI database and the ezbiocloud database (<https://www.ncbi.nlm.nih.gov/genome> &

<https://www.ezbiocloud.net>). From these whole genome data, it is evident that *AP* has a high GC content when compared to other obligate intracellular organisms in the Rickettsiales order at ~41.6% on average (Battilani et al., 2017). There are currently no associated plasmids (Barbet et al., 2013). The genes required for the biosynthesis of lipopolysaccharide and peptidoglycan are also conspicuously missing (Lin & Rikihisa, 2003). Furthermore, *Ap* has a limited coding capacity for central intermediary metabolism, and may only synthesise four amino acids (glycine, glutamine, glutamate, and aspartate); meaning *Ap* must acquire the other amino acids and compounds from host cells (Battilani et al., 2017). Genome data are available (Table 2) with sizes ranging from 1.2 – 2.1mbps, it is worth noting however that most good quality assemblies of *Ap* put that range at around 1.5mbps on average with minimal deviation. Only eight countries have data available from their regions, primarily including: USA, France, and Norway; China, Germany, South Korea, The Netherlands, and Austria each only have one submission.

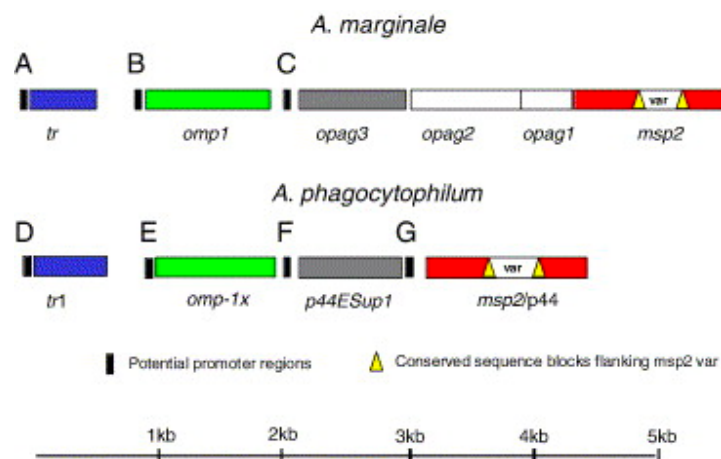


**Table 2: A collection of all publicly available *Anaplasma phagocytophilum* strains sequenced across the world. Source: NCBI GenBank, accessions can be used to access all strains in column 1.**

Accession	Strain	Host	GC%	Size	Fragments	CDS	Country
GCA_000439775.1	JM	Jumping Mouse	41.6	1,481,598	1	1548	USA
GCA_023278635.1	KZ-A1	Human	41.6	1,449,336	1	1514	South Korea
GCA_013487825.1	Norway Variant 1	Sheep	41.7	1,564,418	1	1295	Norway
GCA_000689635.2	Norway Variant 2	Sheep	41.7	1,545,197	1	1627	Norway
GCA_000439755.1	HZ2	Human	41.6	1,447,581	1	1541	USA
GCA_000013125.1	HZ	Human	41.6	1,471,282	1	1558	USA
GCA_000439795.1	Dog2	Dog	41.6	1,473,302	2	1558	USA
GCA_000964685.1	Webster	Human	41.6	1,479,407	1	1579	USA
GCA_000964935.1	HGE2	Human	41.6	1,482,055	1	1563	USA
GCA_000478445.1	HGE1	Human	41.6	1,469,600	2	1571	USA
GCA_000968465.1	HGE1 Mutant	Human	41.6	1,493,666	4	1650	USA
GCA_000964945.1	ApWI1	Human	41.6	1,497,074	1	1597	USA
GCA_000968455.1	CR1007	Eastern Chipmunk	41.7	1,502,024	4	1664	USA
GCA_000965125.1	Annie	Horse	41.8	1,516,523	15	1769	USA
GCA_002849375.1	NY18	Human	41.6	1,387,508	150	1883	USA
GCA_023476575.1	ApMUC09	Horse	41.7	1,520,397	1	1724	Netherlands
GCA_000964725.1	NHC-1	Human	41.6	1,502,749	15	1732	USA
GCA_000964985.1	ApNYW	Human	41.6	1,503,088	16	1693	USA
GCA_000689655.1	MRK	Horse	41.6	1,479,231	9	1711	USA
GCA_000478445.1	CRT38	Ixodes scapularis	41.7	1,506,545	2	1576	USA
GCA_000964915.1	CRT53-1	Ixodes scapularis	41.8	1,570,903	45	1853	USA
GCA_000689615.1	CRT35	Ixodes scapularis	41.6	1,447,016	25	1710	USA
GCA_900078505.1	C1	Cow	41.9	1,682,317	249	2473	France
GCA_900088605.1	C2	Cow	42.2	1,641,350	230	2357	France
GCA_900088625.1	C3	Cow	42.0	1,561,654	191	2232	France
GCA_900088615.1	C4	Cow	42.1	1,604,419	230	2347	France
GCA_900088665.2	C5	Cow	42.9	1,718,615	300	2554	France
GCA_900000025.1	BOV-10_179	Cow	41.5	1,370,818	199	1893	France
GCA_900088655.1	H2	Horse	41.5	2,191,611	580	3451	France
GCA_000964785.1	ApNP	Dog	41.7	1,521,576	1	1843	Austria
GCA_023476575.1	RD1	Roe Deer	42.0	1,585,182	300	2210	Germany
GCA_023476575.1	AKS2020-120P	Himalayan Marmot	41.2	1,261,482	205	1516	China
GCA_900088675.2	H1	Horse	42.0	1,168,754	300	1153	France

Research in the early 2000's identified the major surface antigen 2 (*msp2*) gene as a point of interest in the *Ap* genome (Barbet et al., 2003). The *msp2* gene encodes a ~40kDa outer membrane protein that is capable of triggering antibody responses in the host (Ijdo et al., 1997). *Ap* was found to have mechanisms to vary the presentation of this antigen at a single expression site to avoid immune reactions (Barbet et al., 2003). Therefore, *Ap* is especially good at maintaining chronic

infections; immune responses to antigens may erase the dominant population, but this allows other *Ap* to take over that express a different antigen during the course of infection. This would explain the cycle between undetectable to maximum bacteraemia observed during livestock infections. There are hundreds of paralogues of the *msp2* gene across the species that attempt to ensure the survival of *Ap* in the bloodstream until it is passed to a tick through a blood meal. Using mouse and equine models of HGA, it was found that the *msp2* transcript displayed a high degree of diversity that emerged at 2 days in mice and 7 days in horses which lasts through 21 to 22 days (Scorpio et al., 2008). The rapid and random nature of the *msp2* transcripts suggests that *msp2* recombination is not driven by adaptive or innate immunity and is instead a random and inherent property of the organism (Scorpio et al., 2008).



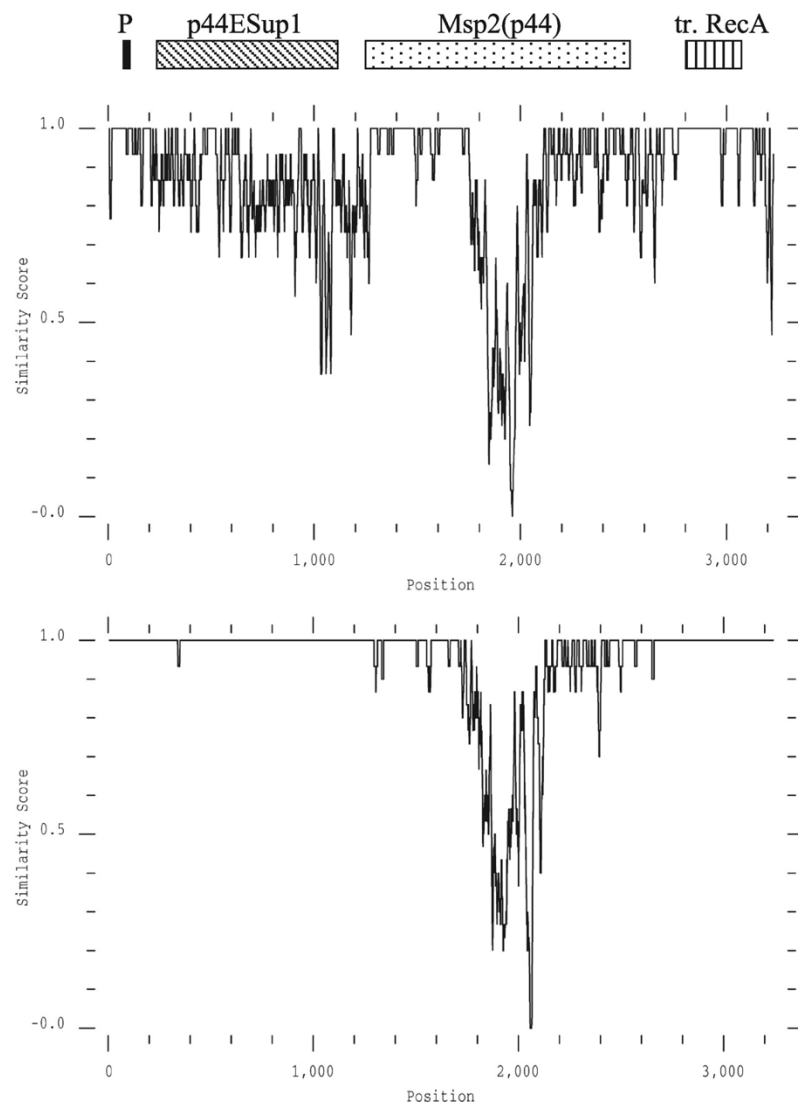
**Figure 7: The *msp2* expression loci of *Anaplasma marginale* and *Anaplasma phagocytophilum*. The black boxes represent potential promoter regions for expression of *msp2* and upstream genes in the loci (Barbet et al., 2005).**

There is a single expression locus within the *Ap* genome, with variable numbers of *msp2* pseudogenes (Barbet et al., 2006). *A. marginale*, a comparable erythrocyte pathogen in the *Anaplasma* genus is a useful model when investigating the *msp2* repertoire in *Ap*. *A. marginale* is generally considered one of the most important tick-borne infections of cattle globally and as a result has been heavily studied, including its mosaic *msp2* expression locus. Figure 7 displays the synteny between *Ap* and *A. marginale*, where *A. marginale* has two extra genes *opag3* and *opag1*. Otherwise, the two species are closely aligned, suggesting the two expression loci may function similarly. *Ap* does however differ by having far more *msp2* pseudogenes than *A. marginale* and as a result does not require segmental recombination of the

pseudogenes to generate sufficient molecular diversity like *A. marginale* (Rejmanek et al., 2012). Over 100 pseudogenes have been observed in a single *Ap* genome some of which may no longer be functional (Rejmanek et al., 2012). Functionality of the pseudogene can be estimated when considering the structure of the gene across the species. All pseudogenes expressed by *Ap* contain highly conserved regions that flank a central hypervariable region (Barbet et al., 2006). The conserved flanking regions have arisen due to the specific folding requirements of the surface protein (Barbet et al., 2006). The orthologs and paralogs of *m*sp2 in different *Anaplasma* and *Ehrlichia* species are unified under Pfam01617. The biological implications of each of these haplotypes of *m*sp2 have not been fully quantified, but future studies that seek to produce complete representations of the crystalised structures could shed light on this.

**NY-18 vs. Swdog  
vs. Norsheep**

**NY-18 vs. NY-37  
vs. HZ**



**Figure 8: *Msp2* expression site variability in three US and two European strains of *Anaplasma phagocytophilum*. A similarity comparison where a score of 1.0 on the Y axis equals identical, with values closer to 0 representing increasing variation (Barbet et al., 2006).**

Interestingly, analysis of the NY-18 strain from the US, showed significantly more homology to other US strains than it did to the European strains Swdog and Norsheep (Figure 8). Thus, suggesting a genetic divide between the continents in terms of surface protein structure despite their close overall phylogenetic ancestry. This is an interesting conclusion as it may be in part contributing to the high number of zoonotic strains in the US.

## 1.8: Conclusion

This chapter explores some of the core concepts and cutting-edge experiments in the fields of *Ap* and *Bartonella* research. Valuable insights have been gained into the diverse genetic landscapes of these microorganisms and their implications for disease, epidemiology and evolution. Crucially, we have evaluated that the *Bartonella* genus is of academic, clinical and ecological importance and although well researched, many gaps still remain in the literature. These primarily relate to the vast number of uncharacterised strains, especially those of rodent origin. Rodent associated strains of *Bartonella* are particularly interesting as they have been proposed as zoonotic pathogens (Krügel et al., 2022). Furthermore, the evolution of the secretion systems within the *Bartonella* genus is debateable. So far, we have established that the systems themselves have come from bacterial conjugation in most cases. Phylogenetic studies have clearly identified links between the presence or absence of certain secretion systems such as the trw-T4SS and flagella, which are mutually exclusive across the genus. However, other systems such as the vbh-T4SS have more mysterious functions and have even been suggested to be not functional altogether. Ultimately it has been identified that the point of acquisition of these systems and the function of some systems is up for debate. The classification of novel species of *Bartonella* will only aid in the elucidation of some of the more complex stories in *Bartonella* evolution and contribute to a better understanding of currently extant species. This thesis will explore the process of genomic, phenotypic and phylogenetic characterisation of a novel species of *Bartonella*, *B. bennettii* and validate the publication of *B. bennettii* as a new member of the genus.

The body of research available for *Ap* is comparatively small, but significant strides have been made to model the evolution and epidemiology of the bacterium. It has been established that *Ap* is a highly diverse species, especially across Europe.

Jahfari et al., 2014 has effectively used the *groEL* gene to delineate European strains of *Ap* into ecotypes, each associated with specific vectors and hosts. The spectrum of genomic diversity of *Ap* has been evaluated using MLST studies identifying discernible clusters linked to geography and host specificity, but with some notable disagreements with the *groEL* based lineages. Complete representations of *Ap* genomes are rare, with only 33 completed to date. Importantly, no *Ap* genomes have been fully sequenced with origins in the UK despite the widespread nature of *Ap* across tick infested regions and the major impacts to livestock farming. Therefore, this thesis will attempt to provide novel insights to the genetic diversity of *Ap* in the UK and Europe. It has also been identified that tick-cell culture, although critically important to research efforts have complicated the process of generating *Ap* genomes. The unculturable nature of obligate intracellular bacteria makes isolation challenging. Consequently, sequencing is less efficient, and specific expertise is needed to generate complete genomes.

---

## 1.9: Aims & Objectives

Both *Ap* and bartonellae strains will be explored in this thesis, with the aim of closing some of the research gaps identified in chapter 1. Firstly, a novel species candidate, *Candidatus Bartonella bennettii* sp. nov. will undergo characterisation according to international code of nomenclature of prokaryotes (ICNP) in chapter 2. This includes complete genomic, biochemical, and phenotypic characterisation of three candidate strains thought to belong to *Candidatus Bartonella bennettii* sp. nov. Chapter 3 will dive into *Ap* genomics in the context of 7 historical UK based isolates maintained in tick cell lines. This chapter will present draft and complete genome assemblies for all seven strains and explore their place within the species complex using a range of contemporary bioinformatical techniques (ANI, phylogenetic placement, gene presence and absence analysis). As no whole genome representations have previously been produced for UK isolates of *Ap*, this work will add to the growing collection of *Ap* genomes, which can aid our understanding of diversity, infection risk and evolutionary trajectory of this important veterinary pathogen. Chapter 4 continues with the exploration of UK *Ap* diversity but sets out to generate complete genomes directly from infected tissues (Blood and Spleen). A comprehensive exploration of enrichment technologies including differential lysis, methylation

depletion, bait capture and adaptive sampling will be individually evaluated and combined to find the optimal methodology for sequencing low population intracellular bacteria from complex samples. Chapters 5 & 6 examine how the information gathered in chapters 3 & 4 can be applied to *Ap* in a global context. This involves a thorough exploration of the widely accepted European ecotypes, including an attempt to sequence a highly divergent shrew derived strain collected in Kielder Forest, UK. Additionally, collaborative datasets provided by Hein Sprong and his research group will be evaluated, including the first human derived *Ap* genome, and the first ecotype III strain of *Ap*. Finally, Chapter 6 focuses on the future directions of this research, evaluating what has worked and what could be improved.

# CHAPTER 2

## 2.0: Characterisation of a Novel Small Mammal-Associated *Bartonella* Species

### 2.1 Introduction

Bacterial taxonomy serves to inform veterinary and public health sectors. As a result, the systems and strategies to define and redefine bacterial taxonomy are critical for disease prevention and control. Technical progress has shifted bacterial classification from overall similarity of phenotypic and genotypic characteristics to more natural theories such as phenetic and phylogenetic relationships between organisms (Vandamme, 2011). Changes to bacterial classification protocols to include genomic data brought about an explosion of new bacterial species and the reorganisations of several genera (Janda, 2018). The *Bartonella* genus was no exception to this and was significantly redefined by Birtles et al., 1995, unifying the genera of *Grahamella* and *Bartonella*. There are currently 39 validly published species and 3 subspecies of *Bartonella* that for the most part can be delineated into four primary lineages. These lineages refer to the different phylogenetic clades identified through genetic and ecological factors. For example, L1 contains species like *B. bacilliformis* and *B. ancashensis*, which are primarily associated with human infections; L2 includes species primarily associated with ruminants such as cattle and deer; L3 and L4 are significantly more diverse lineages and include species that are able to infect a wide variety of mammals including rodents, cats and bats. Each lineage and each species or bartonellae exhibit specific host-adapted traits and has evolved mechanisms that allow the bacterium to thrive in its host environment. These adaptations can be traced across the lineages and include common virulence factors such as flagella and T4SS that perform specialised tasks such as host immune modulation and erythrocyte adhesion. Common strategies are employed to characterise each new species of bartonellae based on the ICNP (Oren et al., 2023).

The ICNP outlines four primary requirements: (1) The selection of a viable reference type strain that has been submitted to at least two internationally recognised culture collections to ensure reproducibility. (2) A valid species description that covers morphological, biochemical, genetic and phenotypic properties to ensure the species is well differentiated from existing taxa. (3) The species must adhere to binomial nomenclature (genus followed by species) and follow Latin or Latinised forms to ensure international consistency. (4) The species description and name must be published in a peer-reviewed journal. These rules ensure that new bacterial species names are stable and globally recognisable. ANI is a category of computational analysis that is often used to define species boundaries (Yoon et al., 2017).

Algorithms often fragment chromosomes, align them, and perform pairwise identity analysis; some ANI tools such as OrthoANI will even employ concepts of orthology to improve calculations. The threshold for classification of a new species has been often quoted at 95 – 96%, providing a cut off for overall genome relatedness (Goris et al., 2007; Kim et al., 2014; Richter & Rossello-Mora, 2009). Phylogenetic placement in the era of whole genome sequencing utilises large portions of relevant genomes to infer evolutionary relationships. A common strategy is the identification, extraction and concatenation of core genes shared across all genomes, followed by maximum-likelihood analysis or Bayesian inference. Despite progress in the molecular characterisation of bacterial genomes, no natural borders have been effectively defined between species, with researchers often relying upon agreed thresholds that are common across all genera.

Although not yet valid members of the genus, there are 15 other *Bartonella* isolates that have publications supporting their placement as novel species of *Bartonella*. These isolates have not been added to *Bartonella* nomenclature as their taxonomic status remains unclear due to not fulfilling the ICNP requirements for species validation. Despite this, these isolates do have completed genomes and accurate descriptions, making them valuable additions to phylogenetic investigations (Table 3). There are ten isolates proposed as L4 species, one as a member of L2 and four that cannot currently be associated with any lineage. *B. apihabitans* and *B. choladocola* have been isolated from the gut of honeybees like *B. apis* and may form a fifth lineage of *Bartonella* species that have substantially larger genomes than the rest of the genus. Interestingly eight of the proposed species are rodent in origin



suggesting we have only scratched the surface of *Bartonella* diversity within global rodent populations. This makes the characterisation of rodent associated strains of *Bartonella* particularly important as novel genetic variants may pose threats to human health and provide insights into the mechanisms and drivers of *Bartonella* evolution.

**Table 3: *Bartonella* species currently without standing in nomenclature.**

Accession	Proposed Species	Strain	Lineage	GC%	Size	Fragments	CDS	Plasmid Size
GCA_000341355.1	<i>Bartonella australis</i>	Aust/NH1	n/a	41.8	1,596,490	1	1367	n/a
GCA_000278215.1	<i>Bartonella rattimassiliensis</i>	15908	4	35.1	2,170,653	12	1761	n/a
GCA_000278275.1	<i>Bartonella tamiae</i>	Th239	n/a	37.8	2,260,792	3	1981	n/a
GCA_000278135.1	<i>Bartonella washoensis</i>	Sb944nv	4	37.4	1,970,822	11	1712	n/a
GCA_900185775.1	<i>Bartonella mastomydis</i>	008	4	38.4	2,045,026	12	1669	n/a
GCA_902162175.1	<i>Bartonella sahelensis</i>	077	4	38.4	2,258,762	131	1907	n/a
GCA_902150025.1	<i>Bartonella massiliensis</i>	OS09	4	37.8	2,277,694	91	1876	n/a
GCA_024297065.1	<i>Bartonella harrusi</i>	117A	4	38.5	2,235,184	1	2078	29,892
GCA_002007565.1	<i>Bartonella choladocola</i>	BBC0122	n/a	45.5	2,907,212	1	2456	n/a
GCA_022559585.1	<i>Bartonella machadoae</i>	46A	4	39.0	2,708,601	1	2544	n/a
GCA_002007485.1	<i>Bartonella apihabitans</i>	BBC0178	n/a	45.5	2,601,172	1	2286	n/a
GCA_902825145.1	<i>Bartonella phoceensis</i>	CIP107707	4	38.5	1,828,078	141	1681	n/a
GCA_903679515.1	<i>Bartonella gabonensis</i>	669	4	38.0	1,971,183	121	1720	n/a
GCA_000278255.1	<i>Bartonella melophagi</i>	K-2C	2	37.0	1,571,225	13	1386	n/a
GCA_033318925.1	<i>Bartonella gliris</i>	GG20g1	4	39.5	2,075,725	2	1863	n/a

In addition to these partially characterised isolates, there are a number of *Bartonella* strains that have not been proposed as novel species. Many of these strains have been identified as variants of current species of *Bartonella*, but some such as AR15-3 and a range of *rochalimae*-like strains may represent novel variants within a poorly characterised lineage of *Bartonella* (L3) (Table 4). These strains are very similar to one another showing minimal variation in GC content, size and CDS. High-quality whole genome sequencing of each strain means all but AR15-3 are complete chromosomal representations of *Bartonella*.

One unclassified variant of *Bartonella* has been encountered several times, being referred to as 'BGA' and '*B. rochalimae*-like' (Telfer et al, 2007; Withenshaw, Devey, Pedersen, & Fenton, 2016; Tokacz et al, 2018). *Bartonella rochalimae* is an L3 species of *Bartonella* and a known pathogen in wild carnivore. (Henn et al.,

2009). Like most species of bartonellae *B. rochalimae* resides within the bloodstream of its mammalian hosts without causing significant disease, allowing for a prolonged

**Table 4: A collection of partially characterised lineage 3 strains from the *Bartonella* genus. NCBI GenBank accession numbers are available in column 1.**

Accession	Lineage 3 Strain	GC%	Size	Fragments	CDS
GCA_002810325.1	1-1C	35.9	1,600,621	1	1421
GCA_002022625.1	11B	36.0	1,579,538	1	1407
GCA_002022645.1	114	35.9	1,571,682	1	1434
GCA_002022485.1	A1379B	35.8	1,541,976	1	1367
GCA_002022545.1	CDC skunk	36.1	1,615,323	1	1461
GCA_002022565.1	Coyote22sub2	35.9	1,561,431	1	1406
GCA_002022585.1	Raccoon60	36.1	1,615,700	1	1481
GCA_002022605.1	JB15	35.3	1,494,018	1	1366
GCA_002022665.1	JB63	35.3	1,493,693	1	1341
GCA_002022445.1	AR15-3	35.9	1,630,082	2	1527

bacteraemia that facilitates transmission through fleas. The bacterium employs T4SSs and other virulence factors such as flagella to manipulate and colonise host cells. These ecological adaptations help to explain the sporadic appearance of *B. rochalimae* in domestic animals and sometimes humans, as spillover events occur when these species come into contact with infected wildlife (Ernst et al., 2020). The L3 clade is home to a diverse collection of isolates but just two validly published species: *B. rochalimae* and *B. clarridgeiae*. This novel *rochalimae*-like strain was first isolated by Telfer et al., 2007 in field voles (*Microtus agrestis*) between 2001 and 2004 in Kielder Forest, Northumberland. Telfer's paper investigated the contrasting dynamics of four *Bartonella* species, including the *rochalimae*-like strain in cyclic populations of field voles and their fleas. Through identification and statistical analysis of these *Bartonella*, several findings were made. This included a positive correlation between field vole and wood mouse populations and the number of *rochalimae*-like isolates encountered. They also found that the strain was commonly isolated from older individuals and was the sole *Bartonella* species to be influenced by changes in flea populations. Several implications can be made from these findings. Firstly, it can be hypothesised that the *rochalimae*-like strain, like other

*Bartonella* species, is typically a chronically infecting parasite that is able to co-exist within its reservoir host without impairing survivability. Furthermore, it can be theorised that it is vectored by one or more rodent flea species. Withenshaw and colleagues (2016) were the next group to encounter the strain whilst investigating the between species transmission of *Bartonella* in bank voles (*Myodes glareolus*) and wood mice (*Apodemus sylvaticus*). The *rochalimae*-like strain was rarely isolated from their captured wood mice, and was absent in bank voles in the UK, suggesting it was host-specific. One of the more recent groups to encounter the strain were Tołkacz and colleagues (2018) when they were investigating the presence and diversity of *Bartonella* in various species of voles in Poland. Again, the strain was rare, only being isolated once from a pregnant common vole (*Microtus arvalis*). However, despite repeated encounter in rodents and an apparent widespread distribution in Europe, no attempt has yet been made to characterise this strain or formally integrate it into *Bartonella* taxonomy.

This chapter reports the polyphasic characterisation of three *rochalimae*-like strains of *Bartonella*, C271, J117 and D105 first isolated in the study by Telfer and colleagues (2007) with the aim of assessing their taxonomic integrity and, if appropriate, formally proposing the creation of new species to accommodate them. The characterisations performed will include whole genome sequence comparisons, thereby allowing exploration of gene content, specifically those encoding putative T4SSs and other virulence factors.

In addition to the characterisation of the *B. rochalimae*-like strain, a second rarely encountered isolate was collected from field voles in Assynt Forest, Scotland which has been previously named and validly published as an L4 species: *B. hiexiaziensis* (Li et al., 2015). This species of bartonellae was originally encountered in the Hiexiazi island area of China but has also been detected in Arctic and subarctic regions (Buhler et al., 2022). The primary reservoir hosts of *B. hiexiaziensis* are thought to be small rodent populations. Despite valid inclusion into *Bartonella* nomenclature, there is no complete chromosomal representation available, limiting our ability to represent the species in whole genome derived molecular analyses. Therefore, I saw it fit to sequence and provide the first chromosomal representation of *B. hiexiaziensis* which will be utilised for subsequent analyses of the genus.

## 2.2: Methods

### Trapping & Isolation of *Bartonella* Colonies (Completed by Others)

*Bartonella* isolates C271, J117 and D105 were isolated from field voles trapped in Kielder Forest (55 °13'N, 2 °33'W), which lies on the eastern edge of the border between England and Scotland between 2001 and 2004. All animals were euthanised humanely (following Home Office schedule 1 procedures) and blood was collected from them by cardiac puncture. Blood was then centrifuged at 4000g for 60 seconds and plasma was removed. The remaining pellet of blood cells was stored - 80°C. *Bartonella* were subsequently isolated from thawed blood cell pellets by inoculation onto Columbia agar containing 10% whole sheep blood. Plates were incubated at 35°C in a 5% CO<sub>2</sub> atmosphere for up to 21 days and checked daily for bacterial growth. Putative bartonellae (based on colonial morphology) were passaged onto fresh plates and growth on these was harvested into cryovials containing brain heart infusion broth plus 10% (v/v) glycerol for long-term storage at - 80°C. *Bartonella* isolate RE21 was isolated on the 21/07/2021 by Laura Mackenzie using the same protocol from the blood of a field vole from Assynt Forest in north-west Scotland.

### Revival of Isolates and Extraction of High-molecular Weight DNA

Frozen *Bartonella* isolates were revived by inoculation onto Columbia agar plus 10% whole horse bloodplates, which were then incubated at 35°C in a 5% CO<sub>2</sub> atmosphere. After 10 days incubation, bacterial colonies were harvested for DNA extraction using the Promega Wizard HMW DNA extraction kit (Product Code: A2920), according to manufacturer's instructions. DNA extraction quality was assessed with the Thermo Scientific NanoDrop One, Invitrogen Qubit 3.0 Fluorometer using the broad range dsDNA assay kit (Product Code: Q32850) and the Agilent TapeStation 2200 running genomic screentape (Product Code: 5067-5365) and reagents (Product Code: 5067-5366).

### PCR-based Identification of *Bartonella* species

Primers 443f (5'-GCTATGTCTGCATTCTATCA-3') (Birtles & Raoult, 1996), and 1137r (5'-AATGCAAAAAGAACAGTAAACA-3') (Norman et al., 1995), were utilised

to amplify and sequence a ~700bp fragment of the citrate synthase gene, *gltA* as previously described (Birtles & Raoult, 1996; Birtles et al., 2002). Amplified products were cleaned up with the Invitrogen ChargeSwitch PCR Clean-Up Kit (Product Code: CS12000) according to manufacturers' instructions, quantified as using the NanoDrop One and adjusted to 10 ng/μL. The samples were then subjected to Sanger sequencing (both strands using the primers described above) using a commercial service (Source Bioscience, Nottingham, UK). Sequence data were combined and verified in the Geneious software package (Genious Prime 2023.2.1 (<https://www.geneious.com>)). MEGA11 was used to align the *gltA* sequences obtained with those of other relevant *Bartonella* species using the muscle tool, ends and gaps were trimmed and the alignment was processed to produce a 100-replicate maximum-likelihood phylogenetic tree.

### Biochemical Characterisation of Isolates

The 20NE API strip kit produced and distributed by Biomerieux (<https://www.biomerieux.com>) for identification of gram negative non-enterobacteria was used to obtain a biochemical profile of isolates C271 RE21. The test was carried out according to manufacturer's instructions with the exception of incubation in a CO<sub>2</sub> enriched (5%) incubator at 37°C for 10 days to ensure the bacterium is able to grow.

### Scanning Electron Microscopy

*Bartonellae* were harvested from Columbia agar + 10% horse blood plates after 3 days growth when colonies were small and scant by placing glass microscope coverslips onto the agar surface then gently removing them. The coverslips were then gently rinsed with PBS to remove any growth media. Bacteria were then fixed onto the coverslip at 4°C overnight with SEM fixative (4% paraformaldehyde (PFA) +2.5% glutaraldehyde (GA) +0.1M cacodylate) provided by the University of Liverpool Shared Research Facility. Coverslips were then rinsed three times using (4%PFA+2.5%GA+0.1M cacodylate), allowing 5 minutes per wash. Samples were post-fixed for 1 hour in 1% osmium tetroxide and then rinsed three times using ddH<sub>2</sub>O for 5 minutes each wash. Coverslips were then cleaned incrementally with 50%, 70%, 80% and 90% ethanol for 10 minutes per concentration. A final 100% ethanol wash was performed three times for 5 minutes each. Samples were then

chemically dried using HMDS, attached to a stub and sputter coated with gold/palladium. Prepared coverslips were imaged at the Scanning Electron Microscopy Shared Research Facility (SEM SRF) at the University of Liverpool.

### Genome Sequencing

Two approaches were used to obtain whole genome sequence data from bartonellae.

Short read sequencing for isolates C271 and J117 was outsourced to microbesNG for Illumina 2 x 250bp paired end sequencing on a NovaSeq platform. For each strain 1ng aliquots of DNA in 100µL of elution buffer were dispatched to microbesNG. The standard short read service was purchased which aimed to provide 30X coverage of samples. Roughly  $10^8$  cells of isolate RE21 were removed from culture plates using a 5µL loop and used to inoculate the inactivation buffer provided by MicrobesNG. The buffer tube was appropriately labelled and sent to MicrobesNG for hybrid sequencing. MicrobesNG carried out a DNA extraction on the provided cells. The hybrid sequencing service aimed to provide 50X long read coverage on R9.4.1 ONT chemistry and 30X illumina coverage as previously described for the standard short read service.

### Oxford Nanopore

Long reads of isolates C271, D105 and J117 were generated in-house using the ONT minION platform. Genomic DNA libraries were prepared with ligation sequencing kit SQK-LSK109 and barcoded with the native barcoding expansion 1-12 (EXP-NBD104) according to manufacturer's instructions. The barcoded libraries were quantified on the qubit 3.0 fluorometer using the dsDNA high sensitivity assay kit and measured on the Agilent tapestation 2200 using genomic DNA screentape and reagents following the manufactures guides. Quantified libraries were then normalised and pooled accordingly. The pooled library was loaded into an R9.4.1 minION flow cell connected to the minION MK1C system and ran on default settings for 12 hours. The onboard minKNOW 22.08.9 software basecalled reads in real-time using the fast-basecalling method, before depositing read data in fastq format.

### Assembly & Annotation

Raw fastq data from microbesNG and the MK1C ONT sequencer were compiled on the University of Salford Biol-2 server for processing. Long read data were concatenated prior to the removal of adapters using the publicly available PoreChop software. Reads were then processed with the Filtlong software package. A minimum read length of 1000 base pairs was set to remove short reads from the long read dataset. Short read data were similarly processed using the short read optimised software package, FastP to both filter and trim reads based on quality and remove sequencing adapters with default parameters. Processed short and long read datasets were assembled with the unicycler assembly pipeline which constructs genomes with the short read data set first before bridging fragments with long reads. Assembled genomes were annotated with three different annotation packages: RASTtk (Brettin et al., 2015), PROKKA (Seemann, 2014) and ggCaller (Horsfield et al., 2023).

### Phylogenetics

Whole genome phylogenetic trees were constructed based on protein sequences of 344 aligned and concatenated single copy core genes totalling 115485 amino acids. Annotations generated from RASTtk were used to identify and compile core genes. The Bacterial and Viral Bioinformatics Research Centres (BV-BRC) (Olson et al., 2023) bacterial genome tree pipeline was employed for generating 500 replicate maximum-likelihood trees. The pipeline takes annotated genomes, identifies single-copy PGFams and processes them with RAXML (Stamatakis, 2015). Phylogenetic trees were generated in newick format and manually processed into figures using the Geneious Prime software package and Inkscape ([www.inkscape.org](http://www.inkscape.org)).

### Pangenome & Average Nucleotide Identity

The *Bartonella* genus pangenome was generated using the pangenome pipeline available in the Anvi'o software package (Eren et al., 2015). Anvi'o was provided with fasta assemblies and ggCaller external gene calls for all *Bartonella* species and strains. A local html session was used to evaluate and annotate pangenome data. PyANI was used to calculate an ANI matrix of all inputted genomes (Pritchard et al., 2016). Results were output as SVGs and visually improved in Inkscape.

### Synteny

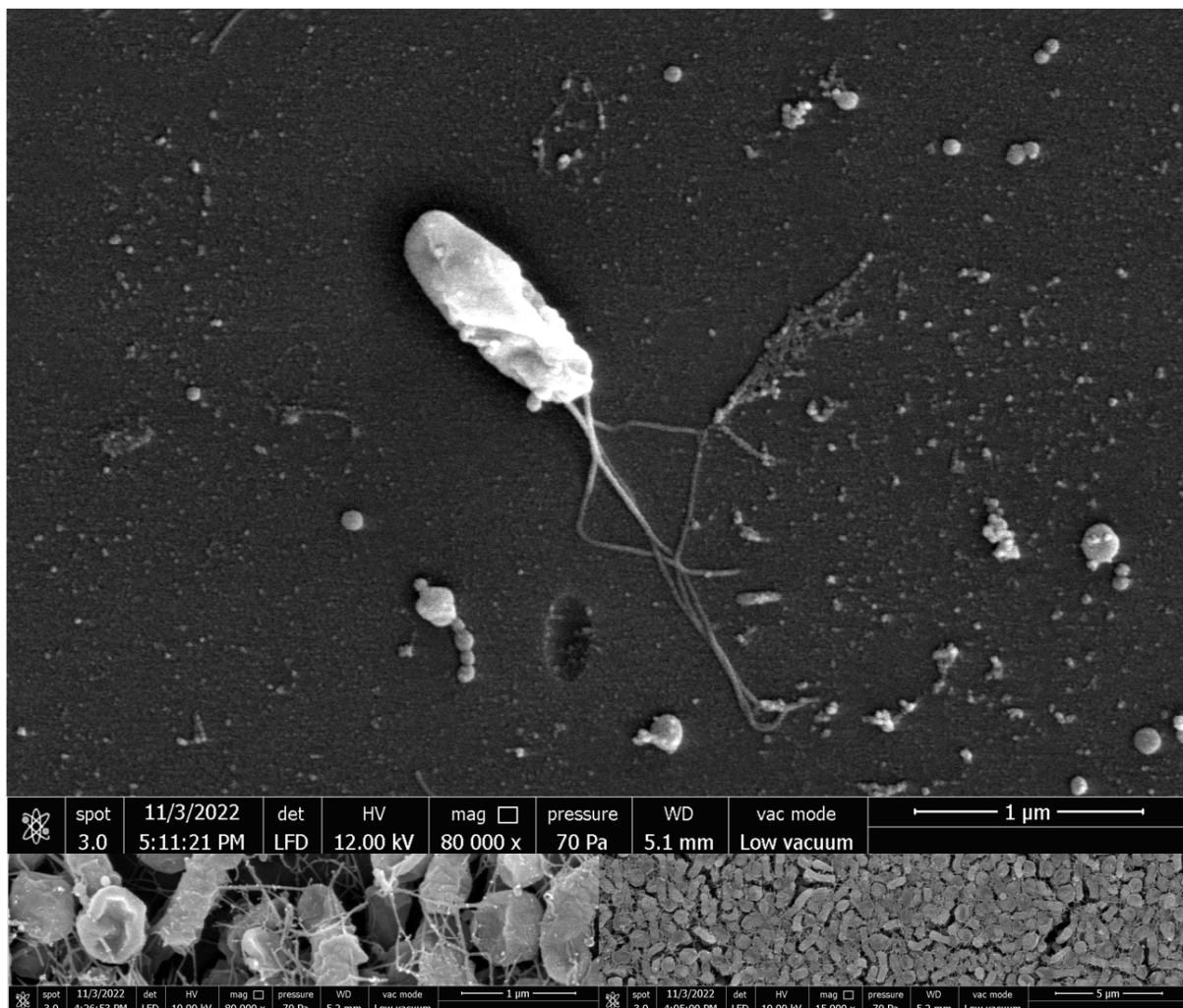


Regions of interest were located and extracted from genomes in the Geneious Prime software package. Extracted regions were saved in GenBank format and inputted into Clinker. Clinker then extracted protein sequences and performed global alignments between sequences, generating images based on cluster similarity. Clustermap.js was used to visually interact with data prior to export (Gilchrist & Chooi et al, 2021).

## 2.3: Results

### Phenotypic Characterisation of *Bartonella* sp. C271

The isolate C271 is a rod-shaped bacillus that measures 0.8µm – 1.2µm in length and 0.3µm – 0.4µm in width. The bacterium possesses polar lophotrichous flagella measuring up to 2µm in length with up to four visible on a single bacterium.





**Figure 9: Scanning electron microscopy (SEM) of *B. bennettii* strain C271 performed at the Liverpool SEM shared research facility. Three images show population structure (15000x), edge of colony (80000x) and an isolated cell displaying polar lophotrichous flagella (80000x).**

### Biochemical Characterisation of *Bartonella* Strain C271

C271 failed to grow in all chambers of the 20NE API strip. Thus, C271 was scored a negative for  $\beta$ -galactosidase (ONPG hydrolysis), decarboxylation of arginine, lysine and ornithine, utilization of citrate, production of hydrogen sulfide, urease, tryptophan deaminase, indole production from tryptophan, acetoin, gelatinase and fermentation of glucose, mannose, inositol, sorbitol, rhamnose, sucrose, melibiose, amygdalin and arabinose.

### Genomic Characterisation of *Bartonella* Strains C271, D105, J117 & RE21

D105 and J117, were identified as additional isolates of the *B. rochalimae*-like strain on the basis of *gltA* sequences which were 100% identical to C271. As a result, D105 (Figure 11) and J117 (Figure 12) were subjected to WGS using a comparable methodology to C271 (Figure 10) utilising both long and short read technology, generating two additional complete genomes for comparative genomic analyses. The *gltA* sequence generated for RE21 implicated the strain as *B. hiexiaziensis* sharing 98.9% of base positions across the 370bp available for comparison. As *B. hiexiaziensis* has not yet been comprehensively sequenced RE21 was also subjected to a WGS workflow (Figure 13). The results of the WGS workflows are summarised in table 5 identifying high homology between the *B. rochalimae*-like strains and a larger genome size of 2.1Mb in RE21 characteristic of an L4 species. Previous attempts to characterise *B. hiexiaziensis* generated three additional gene fragments: *fitZ* (766bp), *rpoB* (775bp) and 16S rRNA (1267bp) enabling further comparisons with RE21. Nucleotide alignments confirmed the identity of RE21 as *B. hiexiaziensis* with 99.7%, 99.2% and 100% pairwise identity respectively.

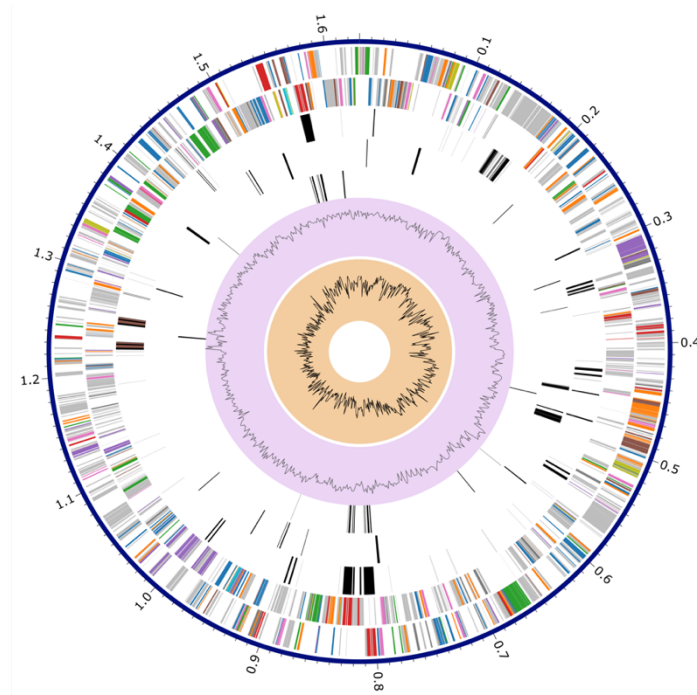


Figure 10: A Circos plot of the *Bartonella* sp. strain C271. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

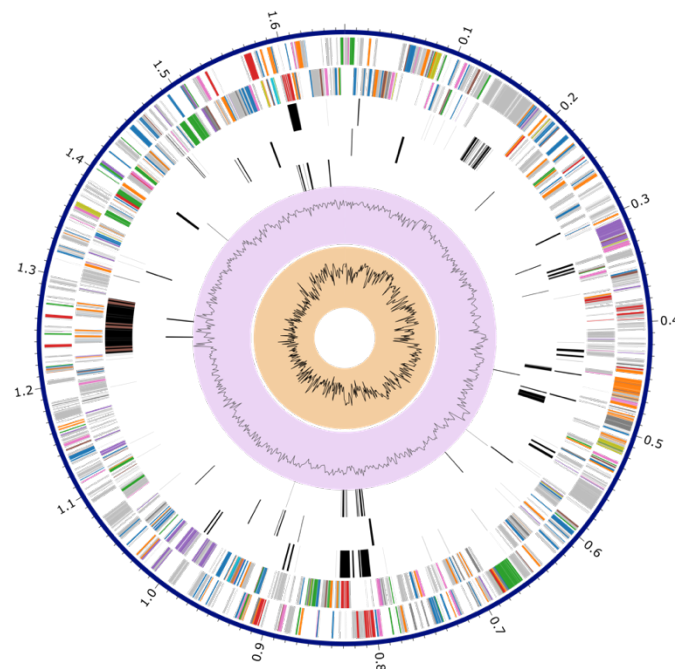


Figure 11: A Circos plot of the *Bartonella* sp. strain D105. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence);

(Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

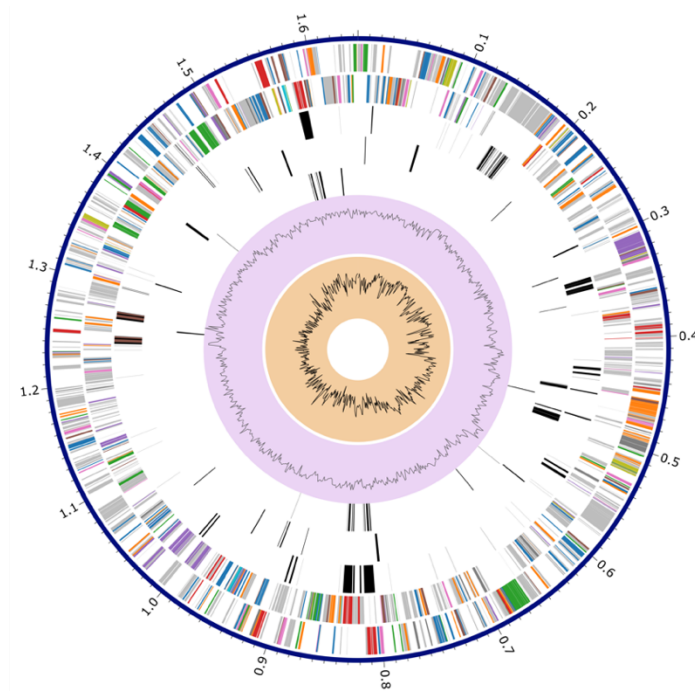


Figure 12: A Circos plot of the *Bartonella* sp. strain J117. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009). Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

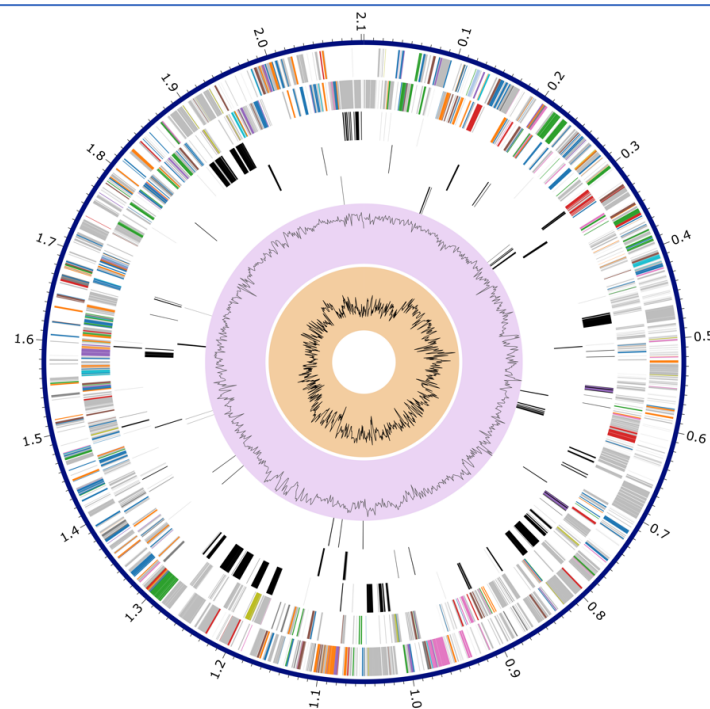


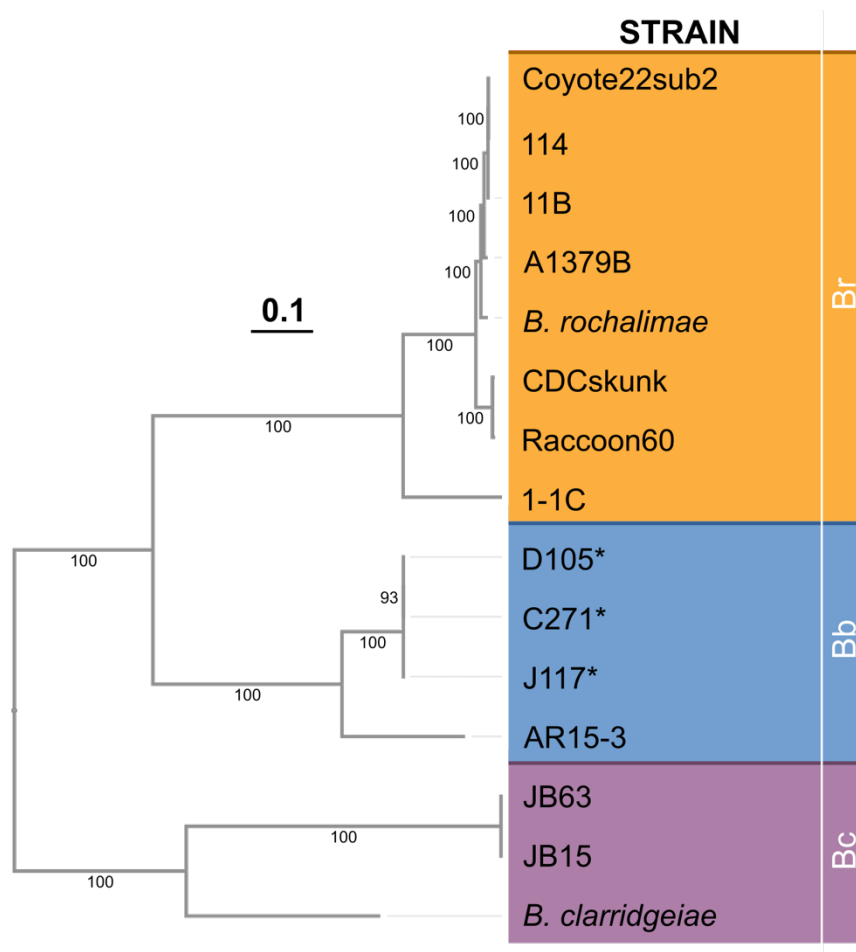
Figure 13: A Circos plot of the *Bartonella* sp. strain RE21. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to

known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew (Krzywinski et al., 2009).

**Table 5: Genome statistics for three strains of *Bartonella bennettii* sp. nov (C271, D105, J177) and *B. hiexiaziensis* strain RE21. CDS annotations were generated with RASTtk.**

Species	Strain	Lineage	GC%	Size	Fragments	CDS
<i>B. bennettii</i>	C271	3	35.9	1,629,655	1	1555
<i>B. bennettii</i>	D105	3	36.0	1,657,696	1	1609
<i>B. bennettii</i>	J117	3	35.9	1,643,736	1	1506
<i>B. hiexiaziensis</i>	RE21	4	39.2	2,102,653	1	2124

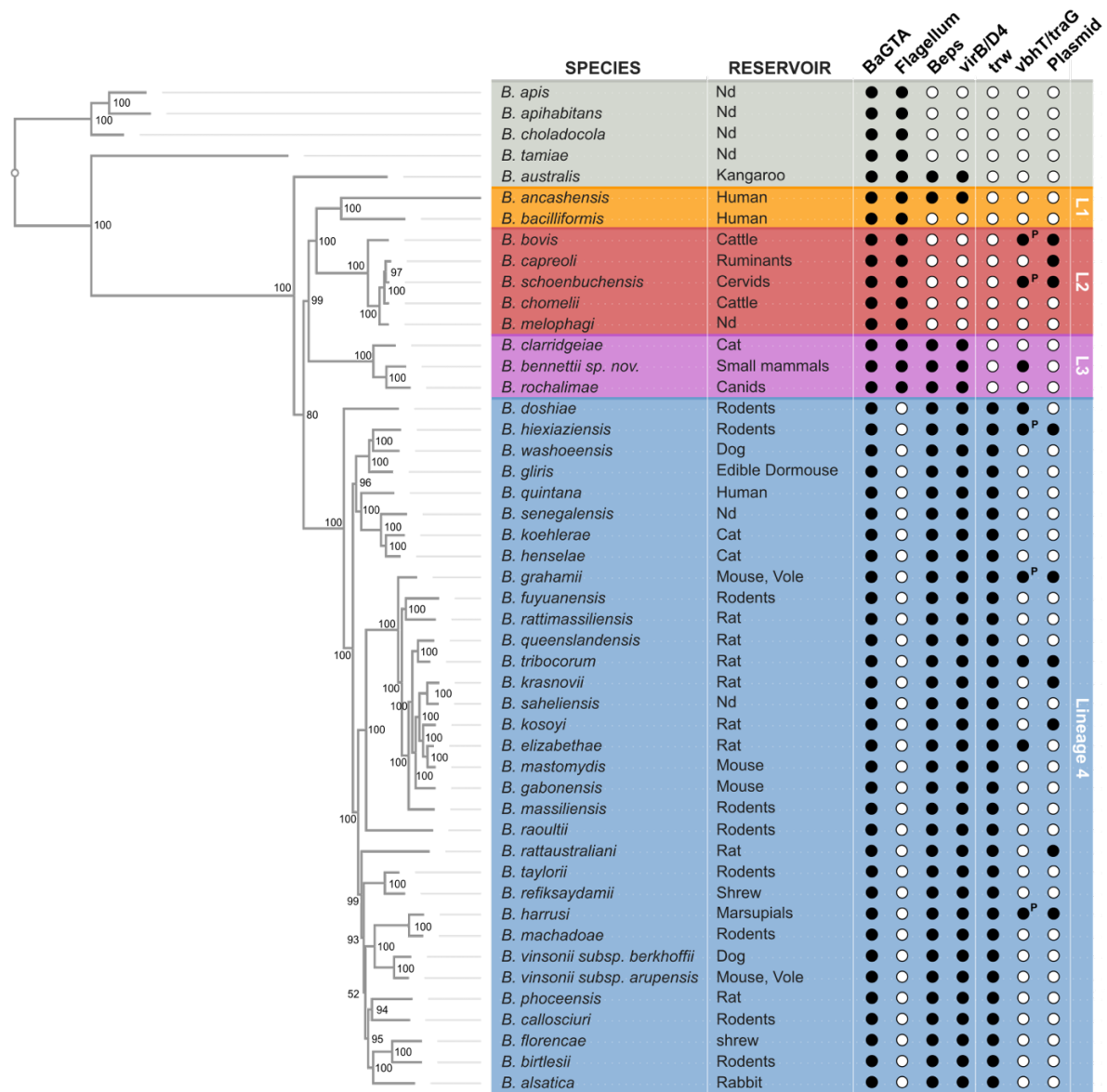
Previous literature and genomic analyses have confirmed the *B. rochalimae*-like strains as divergent members of lineage 3 based on *gltA* sequence analysis, implicating *B. rochalimae* and *B. clarridgeiae* as the closest relatives of the strains that meet ICNP requirements. Comparison of ~300bp fragments of the *gltA* gene



**Figure 14: Phylogenetic tree of Lineage 3 *Bartonella* strains and *B. rochalimae* (Br) and *B. clarridgeiae* (Bc). Strains D105, C271, J117 and AR15-3 are proposed as a novel lineage 3 species *B. bennettii* (Bb). Tree was generated with RAxML using 619,662 amino acid alignments of 500 core single copy genes for each species, bootstrapped with 100 replicates. Strains from this study are marked with \***

with *B. rochalimae* and *B. clarridgeiae* reveal 97.27% and 95.56% pairwise identity respectively.

With the availability of whole genome data for three *B. rochalimae*-like strains (C271, D105, J117), *B. rochalimae*, *B. clarridgeiae* and a number of partially characterised lineage 3 strains outlined in table 4, a whole genome phylogenetic tree based on conserved single copy core genes was constructed (Figure 14). The tree identifies



**Figure 15: Core genome tree of 48 *Bartonella* genomes.** 100 gene maximum-likelihood tree generated from an alignment of 39671 amino acids and bootstrapped with 100 replicates. The presence and absence of key virulence factors is indicated in columns (P) indicates that the vbh1/TraG type 4 secretion system is present on a plasmid.

three clades within lineage 3: *B. rochalimae* group (Orange), *B. clarridgeiae* group (Purple) and *B. rochalimae*-like strain group (Blue). Crucially this phylogenetic tree

provides robust evidence for the integration of strain AR15-3 isolated from an American red squirrel by Inoue et al., 2009 and sequenced by Engel et al., 2011 as an additional *rochalimae*-like strain. I therefore propose the *B. rochalimae*-like strains grouped in this tree: C271, D105, J117 and AR15-3 represent a novel species of bartonellae, *Bartonella bennettii* sp. nov.

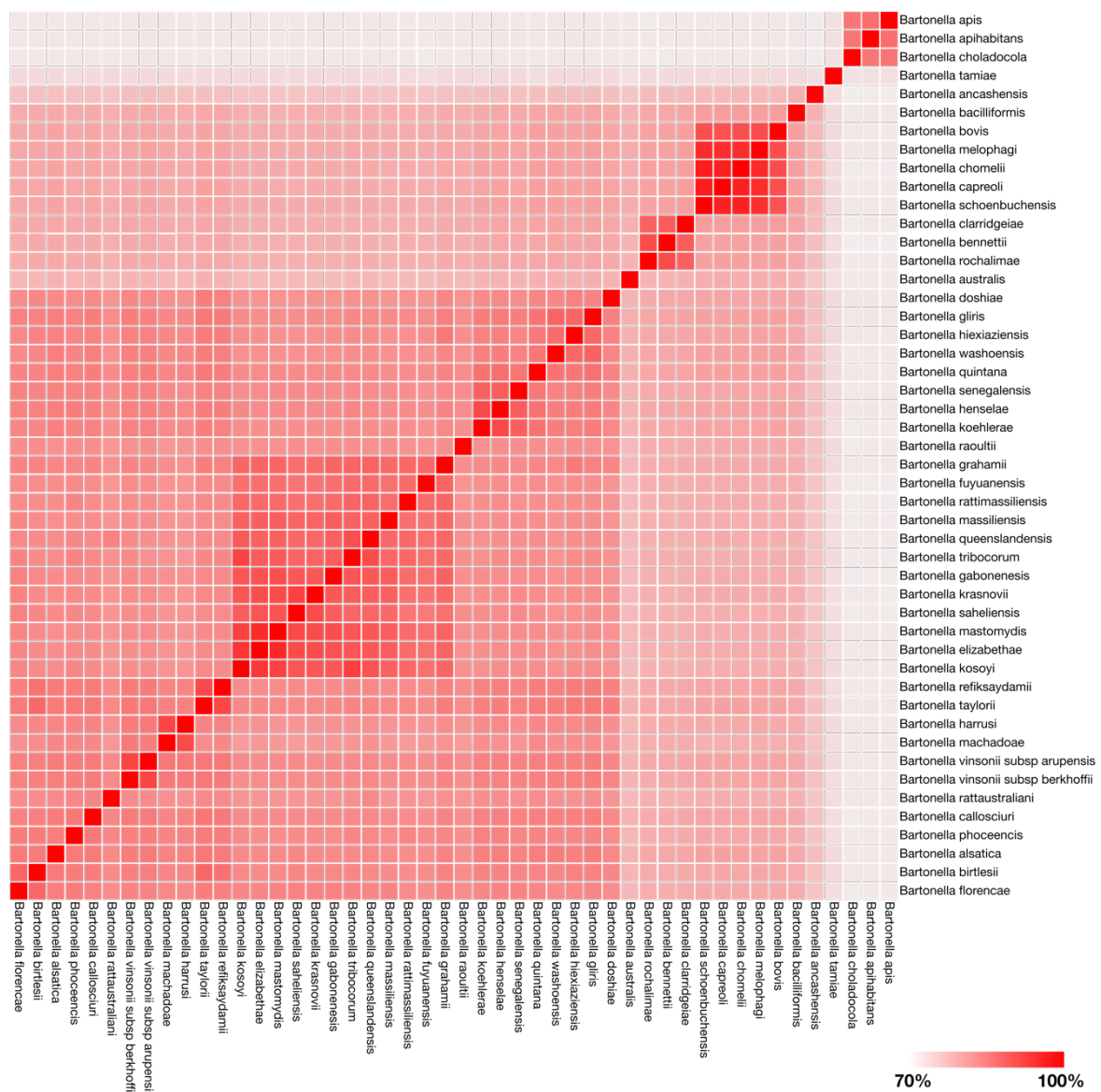
Due to high genetic similarity across the *B. bennettii* strains the proposed type strain C271 was chosen to represent the species in a genus wide phylogenetic analysis (Figure 15). The tree incorporates all validly and invalidly published species of *Bartonella* that have undergone whole genome sequencing including the *B. hiexiaziensis* genome generated as part of this study. This includes 32 of the 39 validly published species in bartonellae nomenclature and the 15 invalidly published species of bartonellae outlined in table 3. Therefore, a total of 48 genomes were analysed via the alignment of protein sequences derived from single copy core genes and processed with RAXML.

In total, 100 single copy genes were used for this analysis, equalling 39671 amino acids. The tree was bootstrapped with the fast-bootstrapping algorithm for 100 replications, generating robust evolutionary predictions across all species. The tree clearly displays the four currently accepted lineages of bartonellae and points to the existence of at least one other lineage comprised of insect symbionts *B. apis*, *B. apihabitans* and *B. choladocola*. *B. bennettii* (C271) fits within L3 with the closest known relative being the canid associated strain *B. rochalimae*. *B. hiexiaziensis* (RE21) resides within L4 within a subcluster containing prominent species such as *B. quintana* and *B. henselae*, two capable human pathogens.

The presence and absence of key virulence factors: the *Bartonella* gene transfer agent (BaGTA) flagella, Beps, and T4SSs were evaluated on both the chromosome and plasmids of the bartonellae species. Only L4 species of *Bartonella* have acquired the *trw* T4SS and lost their flagella. All other species had retained flagella but failed to acquire the secretion system. In keeping with other L4 species *B. hiexiaziensis* possesses the *trw* T4SS, *B. bennettii* on the other hand has not acquired the *trw* T4SS but instead possesses flagella (Figure 15). This is a finding previously noted in chapter 1 whereby the presence of these two virulence factors is mutually exclusive across the genus. The *virB/D4* T4SS appears to have been

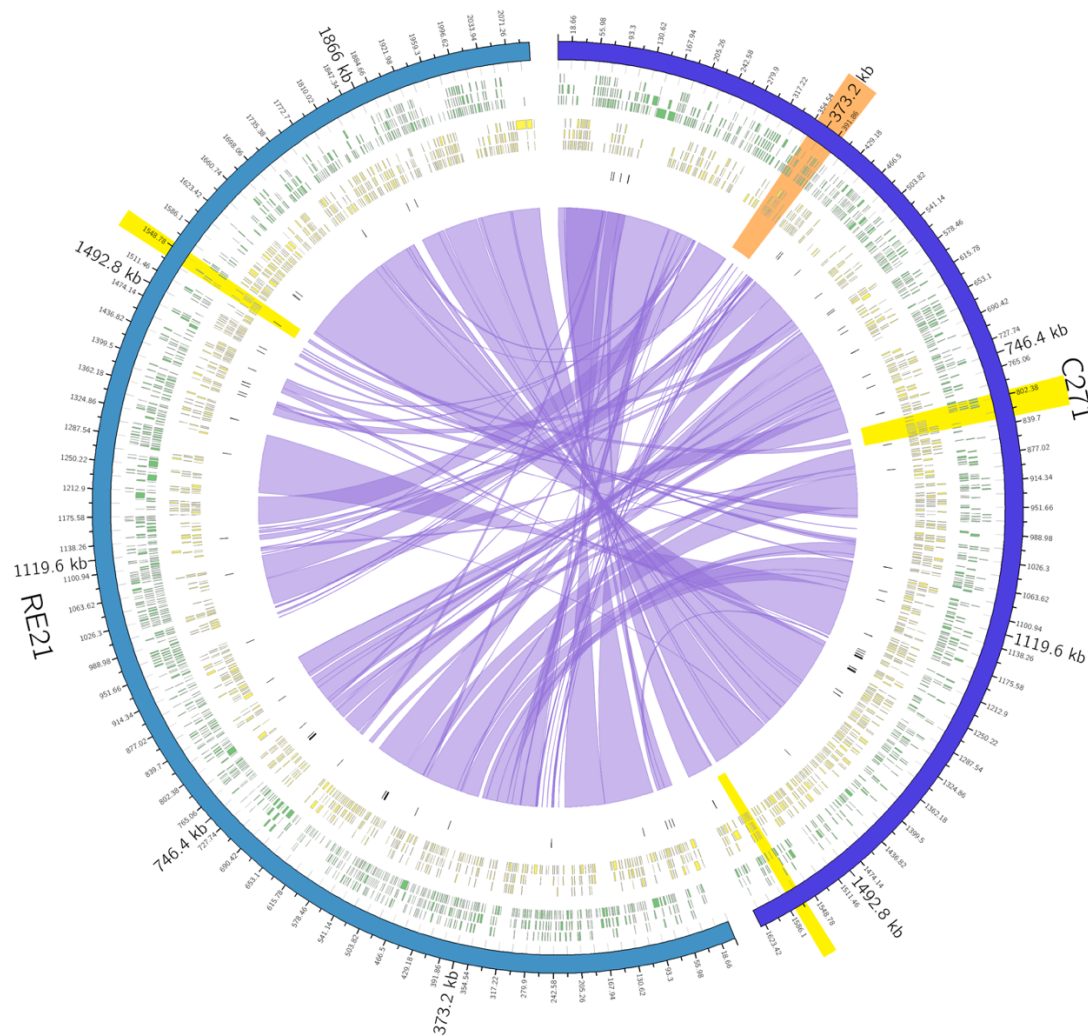


separately acquired through multiple bacterial conjugation events throughout the evolution of the currently extant species of *Bartonella*. The presence and absence of the *vbh/TraG* T4SS is more cryptic, being present in both the chromosome and plasmids of different species with no apparent pattern. Ten species had plasmids available on NCBI GenBank, half of which had copies of the *vbh/TraG* T4SS. L4 was the largest bartonellae lineage being comprised of 33 species. There does appear to be at least 3 subdivisions within L4, but the consequence of these clusters remains elusive. The majority of L4 species have been isolated from rodents, whilst other lineages appear to have an affinity for larger vertebrates such as humans' canines and livestock. *B. bennettii* is a unique addition to L3 having the only copy of a *vbh/TraG* T4SS whilst also being the only species isolated from a rodent.



**Figure 16: Average nucleotide identity heatmap of 48 *Bartonella* genomes calculated with pyANI. Values available in the appendix.**

Calculation of ANI values across the genus indicated that inter-species genome similarity ranged from 70.7% up to 96.5%, with the endosymbionts *B. apis*, *B. choladocola* and *B. apihabitans* displaying the greatest overall variation from other members of the genus (Figure 16). Several members of L2 exceeded the threshold of 95% identity including *B. capreoli* – *B. chomelii* (95.9%), *B. schoenbuchensis* – *B. capreoli* (96.0%), *B. schoenbuchensis* – *B. chomelii* (96.0%). *B. bennettii* C271 is most similar to *B. rochalimae* with an ANI of 90.4% whilst *B. hiexiaziensis* shared the greatest homology with *B. gliris* (87.2%).



**Figure 17: Circos plot illustrating synteny between two *Bartonella* genomes RE21 (blue), C271 (purple) from species *B. hiexiaziensis* and *B. bennettii* respectively. (Orange highlight) likely bacterial conjugation event of a plasmid containing the vbh/TraG T4SS. (Yellow highlights) virB/D4 T4SS locations.**



A Circos plot was constructed to display the synteny between RE21 and C271, more broadly displaying synteny between L3 and L4 species. There is evidence of large recombination events between L3 and L4 as regions of the genomes share high pairwise identity, but significant overall rearrangement (Figure 17). The locations of the Bep secreting virB/D4 T4SS (yellow) are marked on both genomes in addition to a 33kb region on C271 (orange) which contained both a *vbh/TraG* T4SS likely integrated into the chromosome through bacterial conjugation and a *tnpR* gene that can be associated with T3 transposons.

To further explore the virB/D4 T4SSs acquired by the *Bartonella* genus Clinker was used to generate synteny comparisons of extracted operons between validly published species of bartonellae. The focus of these comparisons was to determine whether the novel strains C271, D105, J117 and RE21 had any structural variation when compared to their closest relatives. The first copy of the virB/D4 T4SS is highly conserved across L3 with *B. bennettii* strains showing the highest similarity to *B. clarridgeiae* due to the presence of a *BepA* gene (Figure 18). D105 was the only strain to have internalised stop codons within *virB10*, which resulted in the annotation of 3 fragments of *virB10* gene. The second and third copy of the virB/D4 T4SS have undergone significant rearrangement with rotations of genes and clusters of genes (Figure 19). Again the *B. bennettii* strains share a greater genetic likeness with *B. clarridgeiae* due to the presence of a full third set of *virB* genes (*virB2-11*). *B. rochalimae* was the only species to have just *virB2* and *virB3* make up the third copy of the system. There is just one copy of the virB/D4 T4SS in L4 *Bartonella* species. The system is highly conserved across the lineage but did display significant variation in the number and type of Bep genes present upstream of the operon. RE21 was most similar to *B. henselae* in terms of overall similarity, however all species of L4 shared high homology, especially for the virB/D4 T4SS components displaying no structural variation. The number and orientation of Bep genes did however vary between species, for example RE21 contained eight Beps whilst *B. quintana* contained just five. Interestingly *B. quintana* and *B. hiexiaziensis* were the only species in the comparison to contain a unique Bep, not present in any other bartonellae in the comparison.

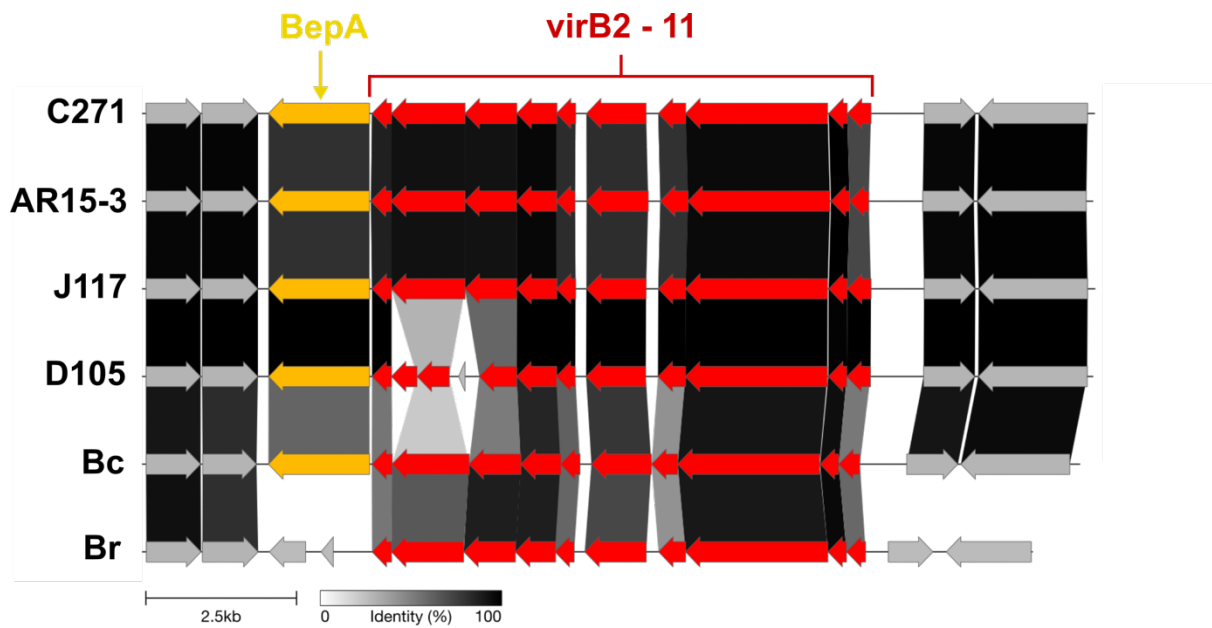


Figure 18: Synteny comparison generated in Clinker of the small virB/D4 type 4 secretion system in lineage 3 species of *Bartonella*. (Br) *Bartonella rochalimae*, (Bc) *Bartonella clarridgeiae*, (C271, D105, J117, AR-15-3) *B. bennettii*.

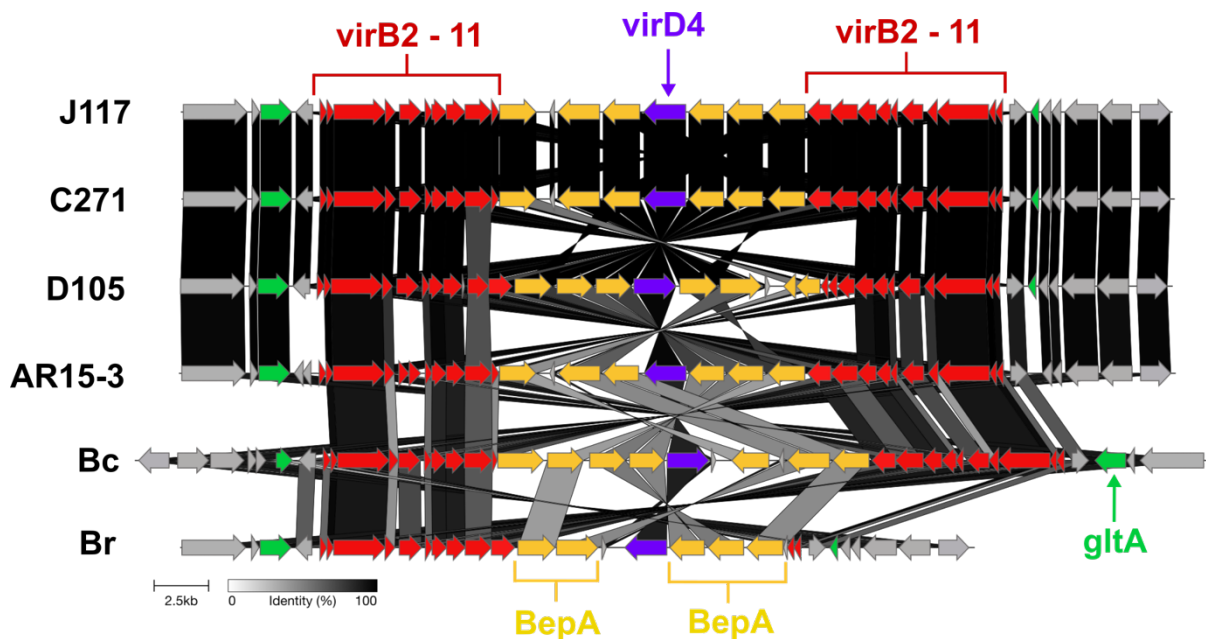


Figure 19: Synteny comparison generated in Clinker of the large virB/D4 type 4 secretion system in lineage 3 species of *Bartonella*. (Br) *Bartonella rochalimae*, (Bc) *Bartonella clarridgeiae*, (C271, D105, J117, AR-15-3) *B. bennettii*.

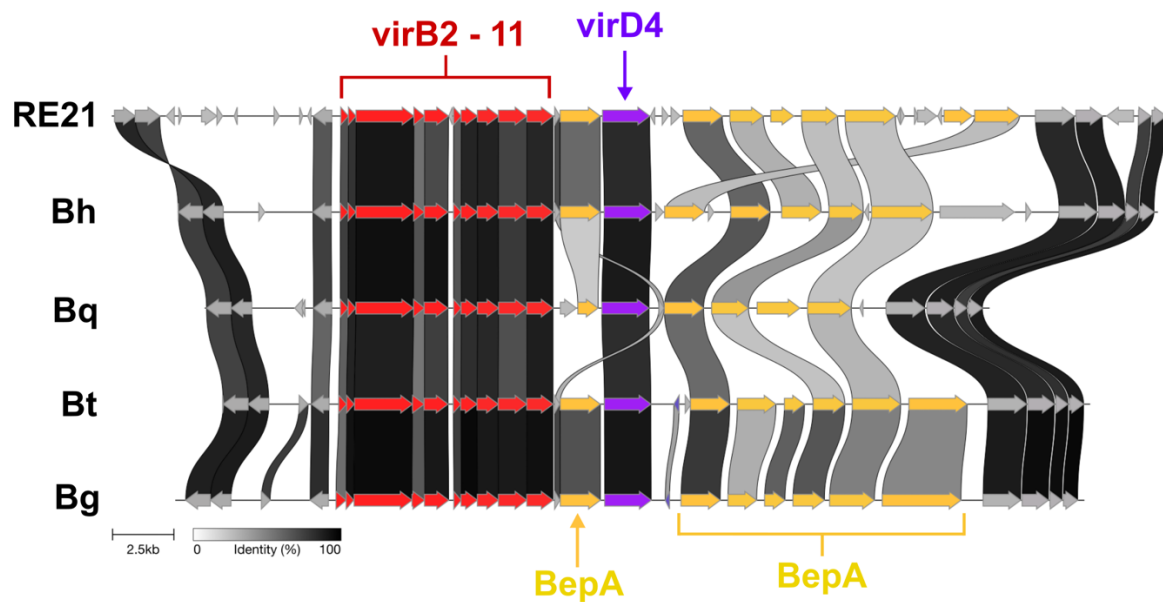


Figure 20: Synteny comparison generated in Clinker of the virB/D4 T4SS in lineage 4 *Bartonella* species, (Bh) *Bartonella henselae*, (Bq) *Bartonella quintana*, (Bt) *Bartonella tribocorum*, (Bg) *Bartonella grahamii*, (RE21) *Bartonella hiexiaziensis*.

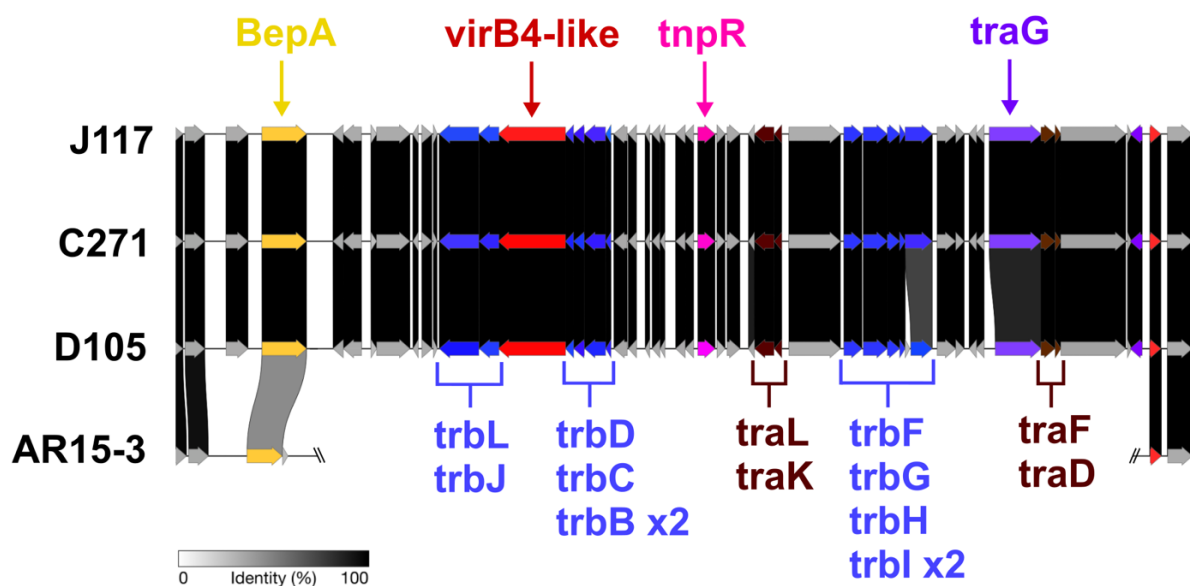
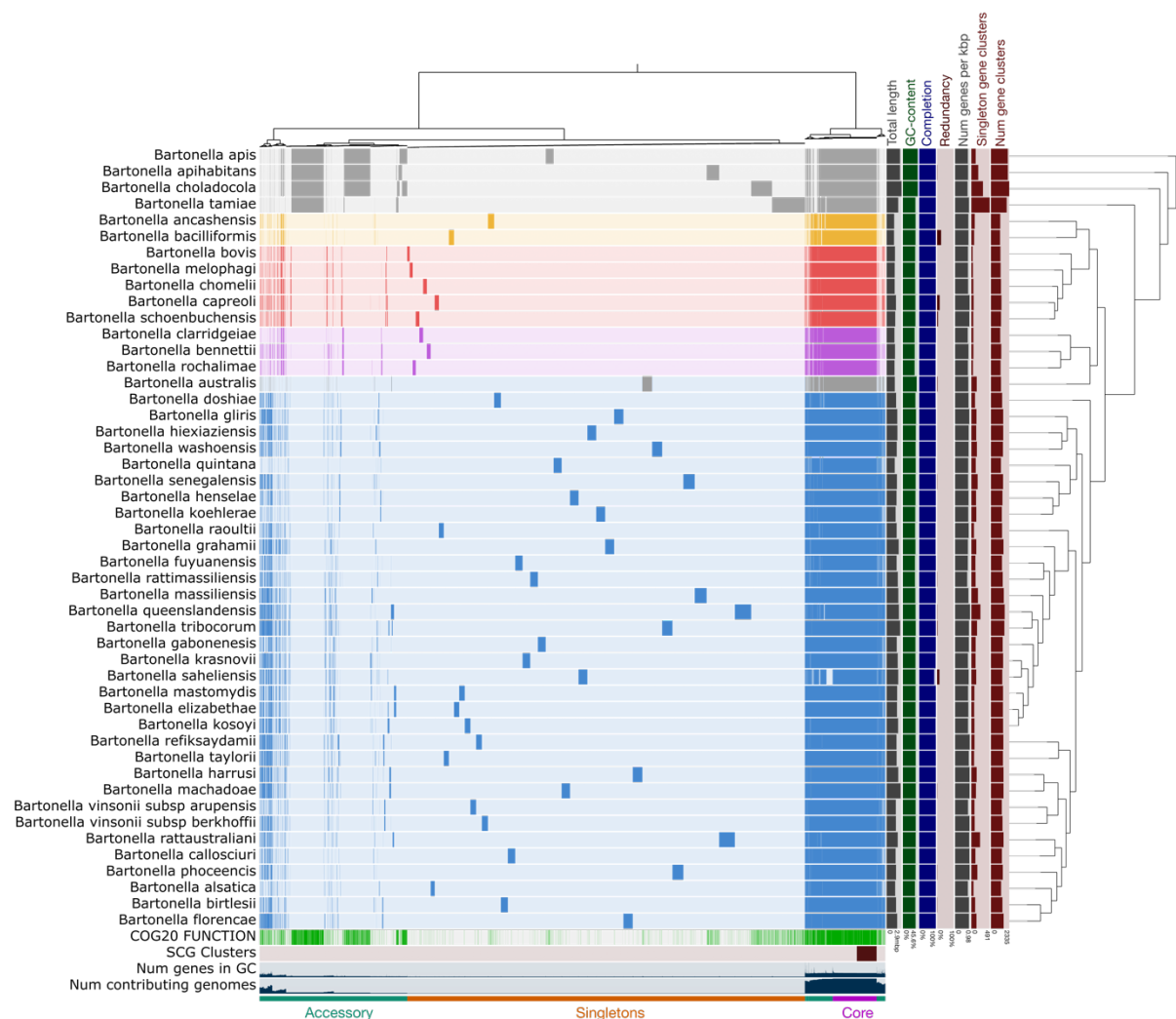


Figure 21: Synteny comparison of the vbh/TraG T4SS in strains of *Bartonella bennettii* sp. nov. generated in Clinker. The System is absent in strain AR15-3.

The vbh/TraG T4SS was present in *B. bennettii* strains C271, D105 and J117 but absent in AR15-3. *B. bennettii* strains C271, D105 and J117 have likely acquired a 33kb plasmid, integrated into the chromosome via bacterial conjugation. There are 48 predicted protein coding genes, 19 of which can be associated with T4SS components listed (Figure 21). Notably, AR15-3 does not present any trace of the

genes associated with the conjugation event in strains C271, D105 and J117 implicating the plasmid may have been very recently acquired and integrated into the chromosome of the UK strains. All strains contained a functional *BepA* gene prior to the predicted conjugation event although AR15-3 had acquired an internal stop codon towards the end of the gene resulting in the annotation of both a smaller *BepA* and a secondary hypothetical protein made up of the missing fragment of the *BepA*. Whether the *vbh*/TraG T4SS is functional in the UK derived *B. bennettii* strains remains unclear, but the presence of key *trb* and *tra* genes may facilitate the expression of a near complete system.



**Figure 22: Pangenome analysis of the *Bartonella* genus generated with Anvi'o displaying 48 species. The layers represent individual genomes organised by phylogenetic relationships based on 295 single copy core genes. Colours represent distinct *Bartonella* lineages (L1 -L4) which are orange, red, purple and blue respectively. Genome completeness, redundancy, GC content, length, genes per kbp singletons, and total number of gene clusters are indicated on the right. On the bottom NCBI COG20 functions (Green line indicates a function was successfully assigned to a gene cluster), number of genomes that contain a specific gene cluster and number of contributing genomes is indicated. In addition to this, the core, accessory and singleton gene clusters are highlighted with purple, green and orange respectively.**

The presence and absence of genes across all 48 genomes was investigated through the generation of a pangenome using the Anvi'o Linux package. Functional annotations were generated with Prodigal (Hyatt et al., 2010) and compiled into databases. There are 9356 orthologous gene clusters that can be categorised into core genome, singletons and accessory genome (Figure 22). Just 295 genes made up the single copy core genome in this analysis. The single copy core genome was used to generate a phylogenetic tree in the Anvi'o package (visible on the right of Figure 22) that was used to order the species into the known lineages. When considering both single-copy and multi-copy core genes there are 657 shared across 48 species which makes up just 7% of gene clusters. The accessory genome made up of 93% of gene clusters in the analysis, which can be split into shared genes and singletons (Figure 22). Of these genes, 2750 can be assigned to multiple *Bartonella* species, making up 29% and 5946 only occur in one species making up the majority of clusters in the analysis at 64%. Despite this, the accessory genome comprises a minority share of genes across most *Bartonella* species, with some notable exceptions being the honeybee gut symbionts, *B. apis*, *B. choladocola* and *B. apihabitans*. Interestingly, large portions of the accessory genome of these species can also be aligned with the human associated species, *B. tamiae*, which may have retained genetic machinery utilised by endosymbiotic *Bartonella*. As a result, the core genomes shared across these four species is much larger.

COG20 functions were assigned to gene clusters using the NCBI COG database embedded within Anvi'o (Figure 22). Green lines indicate a function was successfully assigned to the corresponding gene cluster. A large number of genes were assigned functions with this analysis, including the majority of the core genome. Unfortunately, COG20 functions and categories could not be assigned to the majority of singletons due to their unique amino acid structures. Genome size and GC content are graphically displayed; precise values can be accessed in tables 1 and 3. Estimations for genome completeness and redundancy have also been estimated using the presence and absence of domain-specific single copy core genes. Thus, *B. sahelensis* genome is predicted to be less complete due to the absence of key housekeeping genes we expect to be present in this genome. Most assemblies were predicted to have a genome completeness of >95% and redundancy <5%, indicating good coverage and quality across the genus allowing for robust genus wide

analyses. Finally, the number of genes per kbp and total number of singletons was measured. *B. bennettii* C271 was predicted to be complete, with low genome redundancy. A comparably small number of gene clusters were *B. bennettii* specific, and when mapped back onto the genome largely appear in a chromosomally integrated plasmid (Figure 21).

#### Characterisation of C271 as the type strain of a new *Bartonella* species, *B. bennettii* sp. nov.

The isolate C271 is a rod-shaped bacillus that measures 0.8µm – 1.2µm in length and 0.3µm – 0.4µm in width. The bacterium possesses polar lophotrichous flagella measuring up to 2µm in length with up to four visible on a single bacterium. C271 forms visible white-yellow colonies on blood-enriched Columbia agar within 10 days when grown at 35-37°C with 5% CO<sub>2</sub>. Visible colonies appear round and measure between 0.5 and 2mm in diameter. C271 failed to grow in all chambers of the 20NE API strip. Thus, C271 was scored a negative for β-galactosidase (ONPG hydrolysis), decarboxylation of arginine, lysine and ornithine, utilization of citrate, production of hydrogen sulfide, urease, tryptophan deaminase, indole production from tryptophan, acetoin, gelatinase and fermentation of glucose, mannose, inositol, sorbitol, rhamnose, sucrose, melibiose, amygdalin and arabinose.

C271 contains a single circular chromosome measuring 1.63Mbp in length with a GC content of 35.9%. The closest known relative to *B. bennettii* is *Bartonella rochalimae* (ANI 90.4%).

#### Evaluation of ICNP Requirements

The results above provide a comprehensive description of the strain C271's phenotypic, biochemical and genotypic features. Genome analysis including the construction of whole genome phylogenetic trees and ANI provide strong evidence for the characterisation of C271 as a novel species of bartonellae *B. bennettii* sp. nov. These results meet both requirement (2) and (3) of the ICNP for valid publication of a novel bacterial species. Therefore, following submission of the type strain C271 to two internationally recognised culture collections and the publication of this data in a relevant journal, all four ICNP requirements will be met.

## 2.4: Discussion

The goal of microbial taxonomy is to devise a process of classification and identification that is stable, objective and readily available. (Yoon et al., 2017; Kundisch et al., 2021). In most cases genomic metrics are reliable objective datasets that have improved predictions of bacterial taxonomy (Hayashi Sant-Anna et al., 2019). It is however critical that we evaluate novel algorithms for calculating genome-based relationships between bacteria and layer outputs with the goal of generating more reliable conclusions. (Yoon et al., 2017; Kundisch et al., 2021). Additionally, the data being used to evaluate bacteria at the species level must be scrutinised, with sequencing methodology, coverage and depth being some of the key metrics that can drastically influence outcomes (Chun et al., 2018). This is especially important when genomes are incomplete and fragmented, which is commonplace in metagenomic studies (Meziti et al., 2021). Although utilisation of genomic methodologies remains an optional step in the classification of novel species it is now highly recommended by journals and will likely become mandatory in the near future. This reiterates a need for standardised, readily available methodologies that can be applied by researchers with limited experience in bacterial genomics (Chun et al., 2018). In the case of the *Bartonella* genus, genome completeness is high, and redundancy is low, enabling the application of high-resolution analyses illustrated throughout this chapter.

High-throughput sequencing technologies generate an astonishing amount of information that can be used to generate complete de novo assemblies of bacterial genomes which may elucidate mutations and structural variants that distinguish organisms (Alkan et al., 2011). The de novo assembly process consists of grouping reads based on base identity to represent the complete genome, a process that is often difficult to resolve in cases of highly repetitive elements, areas of low coverage due to biases such as GC bias, and sequencing artifacts peculiar to each platform (Chen et al., 2013; Ross et al., 2013). The sequencing and de novo assembly of *Bartonella bennettii* D105 provides insights into the advantages and disadvantages of commonly employed sequencing equipment. For example, initial attempts to assemble the genome with Illumina short reads failed to produce a complete representation of the genome but rather a draft sequence made up of 338 contigs

with a total size of 1.55mbp. Crucially, key genetic components such as the second and third copy of the *virB/D4* T4SS were absent from this assembly and the high number of contigs complicated genome annotations, generally favouring an increased number of CDS when compared to later annotations of the complete assembly. Long read sequencing platforms such as PacBio and ONT provide solutions to the shortfalls of short read sequencing systems as reads are no longer orders of magnitude smaller than the genomes they represent. In the case of D105 ~100x coverage of the genome with ONTs minION was sufficient to correct the genome assembly increasing total size to 1.66mbp across a single contig. Importantly, ONT reads alone were capable of generating a contiguous genome assembly but lacked the accuracy of reads generated with Illumina platforms. Therefore, the final assembly was generated using a combination of both long and short reads, which is known as hybrid assembly, and widely considered the gold standard for generating de novo genomes (Zhang et al., 2021). We opted for a largely automated assembly package, unicycler, which prioritises assembly with Illumina reads to preserve basecall accuracy, with ONT long reads later mapped onto the Illumina assembly graph to bridge gaps and resolve the genome (Wick et al., 2017). Due to the success of this assembly methodology all future sequencing followed this procedure, resulting in the complete assembly of four novel *Bartonella* genomes.

Logically, the next step was the annotation of genomes to identify functional elements along the sequence, thus giving meaning to the data. Genome annotation is error prone when assembly quality is low as assembly errors may compromise the inference of the true gene function due to reduced similarity with database genes, leading to an increased number of hypothetical proteins. Several annotation packages were evaluated including Prokka (Seemann, 2014), RASTtk (Brettin et al., 2015), and ggCaller (Horsfield et al., 2023), each producing similar results. These annotation packages employ an *in-silico* approach that searches for signal sensors (TATA box, start and stop codons, or poly-A signal detection), content sensors (GC content, codon usage, or dicodon frequency detection), and similarity detection (between proteins from closely related organisms, mRNA from the same organism, or even reference genomes when available) (Stein, 2001; de Sa et al., 2018). Deciding on the annotation package is an important but often partly subjective step,



in the case of the *Bartonella* genus RASTtk was sufficient for identification of protein coding genes, tRNA, rRNA and structural repeats.

In the era of whole genome sequencing, single gene studies often take a backseat when evaluating strain and species diversity. There does however remain great utility in single gene analyses when phylogenetically representative genes are chosen (Guterres et al., 2019). In this perspective, *rpoB* and *gltA* were found to be the most potent genes for analysis of *Bartonella* (La Scola et al., 2003) and as such have been the subject of countless studies of diversity and epidemiology (Lin et al., 2012; Han et al., 2017; Olival et al., 2015). A ~700bp fragment of the *gltA* gene therefore provides ample information for determining strain identity and novelty and has been used to identify strains from large spatial studies where whole genome sequencing is not practical (Morway et al., 2008; Bai et al., 2015). Phylogenetic analysis of *gltA* fragments clearly delineated species of *Bartonella*, identifying AR15-3, D105, J117 and C271 as strains of *B. bennettii* and RE21 as *B. hiexiaziensis*. This not only saves on time and costs but also opens up comparisons to a far larger dataset of strains that haven't undergone whole genome sequencing. Nevertheless, whole genome sequencing provides superior resolution and reliability when estimating phylogenetic relationships (Gonçalves-Oliveira et al., 2023).

To further explore *B. bennettii*, two phylogenetic trees were constructed to investigate diversity across the genus and within L3 where *B. bennettii* was predicted to reside. The L3 tree incorporated fifteen strains of *Bartonella* which were analysed using RAxML and 500 single copy core genes shared by all of the genomes. Three clusters were elucidated, which can be assigned to three species, *B. rochalimae*, *B. clarridgeiae* and *B. bennettii*. Importantly, the whole genome analysis of these strains further implicated AR15-3 as a strain of *B. bennettii*, showing minimal evolutionary distance from the primary cluster of strains (C271, D105 and J117). Interestingly, AR15-3 was isolated from the American red squirrel (*Tamiasciurus hudsonicus*) suggesting *B. bennettii* may be a globally distributed strain capable of infecting multiple species of rodent (Engel et al., 2011).

When considering all members of the *Bartonella* genus, L3 *Bartonella* make up a small portion of the total diversity, residing within the parasitic *Bartonella* cluster that contains all four *Bartonella* lineages that have diverged from their endosymbiotic

ancestors. Amongst Alphaproteobacteria, two molecular machines for transfer of bacterial DNA have emerged, aptly named RcGTA and BaGTA due to their origins in *Rhodobacter capsulatus* and the *Bartonella* genus (Bardy et al., 2020). These gene transfer agents are thought to have evolved from domesticated bacteriophages (Lang et al., 2017). RcGTA is widely distributed across Alphaproteobacteria, whereas BaGTA is restricted to the *Bartonella* genus (Tamarit et al., 2018). BaGTA particles primarily differ from RcGTA particles due to the higher fraction of genes implicated in host interactions (Berglund et al., 2009; Guy et al., 2013). All 48 genomes contained BaGTA, suggesting the component is critical to success with strong selective pressures favouring the functionality of the system. Flagella on the other hand were present in all genomes with the exception of L4 species. As previously mentioned, flagella provide a mechanical advantage to erythrocyte invasion through increased motility, an advantage that appears to be irrelevant when the highly effective trw T4SS is present (Deng et al., 2012; Deng et al., 2010). My analysis corroborates previous findings where flagella and the trw T4SS are mutually exclusive across the genus, implicating the trw T4SS as a superior mechanism for erythrocyte adhesion. It is likely that the trw T4SS was acquired in a single event, prior to the adaptive radiation of L4, implicating the system as a key driver of the success of L4 species.

The virB/D4 T4SS differs from the trw T4SS due to its presence across L1, L3 and L4 *Bartonella* but not all species. All members of L3 and L4 have acquired functional copies with just *B. australis* and *B. ancashensis* having acquired the system outside of L3/4. This suggests that the system has been acquired up to four times across the genus or has been lost by certain species that may not require the advantages provided. The virB/D4 T4SS functions to translocate effector proteins (Beps) into host cells, subverting their functions, aiding in immune evasion (Engel et al., 2011; Wagner & Dehio, 2019). This provides a clear functional advantage to parasitic species expressing the system (Wagner & Dehio, 2019). For this reason, it is likely that several separate acquisitions of the virB/D4 T4SS are responsible for the irregular presence across the 48-genomes analysed. The presence of functional Beps across all of the genomes that have acquired the system further supports this hypothesis. Finally, the presence and absence of the vbh/*TraG* T4SS was plotted. Due to the largely cryptic nature of this system, the advantages of acquiring the

locus remain a mystery (Wagner & Dehio, 2019). Unlike other T4SSs, the *vbh/TraG* T4SS may be present in both the chromosome and the plasmids of *Bartonella* species. Only 9 of the 48 genomes interrogated had evidence of the locus, with just four having a chromosomally integrated copy. *B. bennettii* was one of the four species to have chromosomally integrated the T4SS a curious finding when the highly similar AR15-3 genome has no such copy. Additionally, *B. hiexiaziensis* has also acquired the *vbh/TraG* T4SS within a plasmid sequenced alongside its genome. The *vbh/TraG* T4SS shares high homology with the *virB/D4* T4SS and has been implicated as a classical interbacterial conjugation system that lacks a clear biological role (Harms et al., 2015; Harms et al., 2017; Wagner & Dehio, 2019). Both *B. hiexiaziensis* and *B. bennettii* contain *TraG* genes encoding the crucial accessory component for conjugation, suggesting a functional system may be present (Harms et al., 2017; Wagner & Dehio, 2019).

L3 and L4 *Bartonella* species have acquired the *virB/D4* T4SS in all constituent species. The number of copies of this system does however vary with L3 species holding up to three complete copies, whereas L4 have just one. Gene duplication and indeed whole operon duplications are common in prokaryotic organisms, facilitating rapid evolution of molecular machinery whilst increasing genomic complexity (Fondi et al., 2009; Engel et al., 2011). Figure 17 illustrates whole genome synteny between L4 and L3 *Bartonella*, with *B. hiexiaziensis* RE21 and *B. bennettii* C271 representing each lineage. It is evident from this figure that *Bartonella* genomes have undergone some rearrange of their genome structure but retain their overall genomic backbone. Typically, intracellular bacteria have lower recombination rates, likely due to their relative niche isolation (Vos & Didelot, 2009). The *Bartonella* genus is no exception to this boasting one of the lowest recombination rates amongst all bacterial species (Vos & Didelot, 2009). Rodent associated strains of *Bartonella* do however appear to undergo more frequent recombination events than other *Bartonella* species suggesting a broader host range for rodent adapted species (Berglund et al., 2009; Paziewska et al., 2011). The first copy of the *virB/D4* T4SS in L3 and the only copy of the system in L4 is well conserved with minimal variation across the species and strains investigated. Substantial genetic rearrangement only becomes evident in the second and third copy of the *virB/D4* T4SS in L3 species where a fragment of the *gltA* gene implies the duplication of the operon. Subsequent

rearrangements of the two copies could be the product of recombinases, transposases or integrases (Engel et al., 2011), but no genetic evidence of this appears in the genomes. Differences between operons does however appear to be tied to gene amplification and diversification, a process that has generated the huge variety of Beps in extant species from a single primordial effector (Schulein et al., 2005). One of the major findings in the *B. bennettii* genomes is the presence of the *vbh/TraG* T4SS. Due to the high homology with *vbh/TraG* carrying plasmids, we can assume the acquisition of this system originates from bacterial conjugation of a plasmid. The presence of the *tnpR* gene is however curious, as this implicates transposases as potential culprits of genome rearrangement within *Bartonella*. The absence of the *vbh/TraG* T4SS in AR15-3 provides a unique opportunity to estimate the size of the plasmid integrated into the genome (33kb) (Figure 21).

The final concept explored in this chapter was the pangenome, a collective genetic system of all living organisms (Tettelin & Medini, 2020). The concept alerts to the inherent flow and transmission of DNA between organisms, including horizontal gene transfer between related and unrelated microorganisms (Tettelin & Medini, 2020). Pangenomics allows us to model enormous datasets such as those displayed in figure 22 and draw conclusions from the patterns that emerge. Over 84000 genes across 9356 orthologous gene clusters are investigated, with the core and accessory genomes highlighted. As more organisms are added to an analysis the size of the accessory genome can increase relative to the core genome. A phylogenetic tree calculated using single copy core genes organised species on the pangenome and a number of key genome metrics are displayed (Figure 22). A criterion for assessing genome quality, completeness, and redundancy was utilised to qualify predictions of phylogenetic placement, epidemiology and ultimately taxonomy. An example of the utility of these metrics is evident in *B. tamiae*. *B. tamiae* contains a high number of singletons which may suggest rapid evolution and acquisition of genetic components in response to selective pressures, or poor overall genome quality such as too many contigs, low sequencing depth and/or coverage. To determine whether the former or latter is more likely, prior information such as assembly quality, phylogenetic placement, epidemiology and CDS per kbp and more must be considered. *B. tamiae* has acquired the most singletons in the genus, has good overall sequencing quality (3 contigs across 2.2mbp, >95% completeness, <5% redundancy) with a moderate

ratio of CDS to genome size (1981 CDS across the genome). Furthermore, *B. tamiae* is shown as a phylogenetically diverse species with no associated lineage. Therefore, it is likely that the singleton gene calls within this assembly are true, representing unique genetic machinery evolved to provide a survival advantage to *B. tamiae* in its reservoir hosts and vectors. The same can be said for *B. bennettii* and *B. hiexiaziensis* which have comparably small singleton genomes.

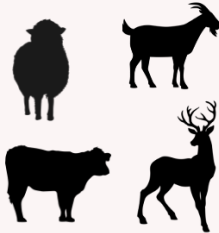



This chapter verifies *B. bennettii* as a novel member of the *Bartonella* genus, residing in L3 alongside *B. rochalimae* and *B. clarridgeiae*. Phenotypic, biochemical and genomic characterisations of three strains, C271, D105 and J117 are provided. Critical points of debate in the microbiological research community are explored such as the species concept and the challenges associated with classifying prokaryotic organisms.

# CHAPTER 3

## 3.0: Exploring the Genomic Diversity of UK *Anaplasma phagocytophilum* Isolates.

### 3.1: Introduction

*Anaplasma phagocytophilum* (*Ap*) is an emerging globally distributed tick-borne parasite of the granulocytes that is capable of infecting a wide range of wild and domestic animals (Dugat et al., 2015). *Ap* causes tick-borne fever (TBF) in domestic ruminants and granulocytic anaplasmosis (GA) in people, equines and canines (Apaa et al., 2023). Human GA is rare in Europe and absent in the UK with few cases diagnosed to date thought to party due to reduced awareness and surveillance (Stuen et al., 2013). TBF on the other hand is widespread impacting animal welfare and productivity on farms situated in areas with high tick burdens (Dugat et al., 2015; Bianchessi et al., 2013). The problems associated with TBF are for the most part experienced by young animals, and individuals reared in tick-free areas that are placed on tick-infested pastures for the first time (Stuen et al., 2013). Characteristic symptoms include high fever, anorexia, dullness, and a sudden drop in milk yields (Tuomi, 1967). Due to the immunosuppressive effects of *Ap*, secondary infections may also point to TBF in herds of animals, with up to 30% of lambs (*Ovis aries*) infected with *Ap* developing tick pyaemia (TP) due to infection with *Staphylococcus aureus* (Woldehiwet, 2006). TP is considered the most common and serious complication of *Ap* infection in UK livestock with some studies estimating that 300,000 lambs develop TP annually as a result of *Ap* infection (Brodie et al., 1986). Most lambs that develop TP die, whilst abortions and reduced milk production in cattle herds further contribute to economic losses (Brodie et al., 1986). Even though data regarding the financial impact of *Ap* in the UK is limited and somewhat out of date, it is commonly accepted that *Ap* represents a significant challenge for UK farmers located in areas of high tick prevalence (Stuen et al., 2013). It is therefore essential that we characterise the strains of *Ap* circulating in the UK in order to develop more specific management strategies, and accurate infection risk maps.

Ecotype I	Ecotype II	Ecotype III	Ecotype IV
<b>Generalist</b> <i>Ixodes ricinus</i> 	<b>Roe Deer Specialist</b> <i>Ixodes ricinus</i> 	<b>Rodent Specialist</b> <i>Ixodes trianguliceps</i> 	<b>Bird Specialist</b> <i>Ixodes frontalis</i> 

**Figure 23: The four ecotypes of *Anaplasma phagocytophilum* detected in European wildlife, their respective host tropisms and vectors.**

Many species of *Ixodes* tick are involved in the transmission and maintenance of *Ap* in populations of wildlife, with only a small number of additional vectors potentially involved such as deer keds (*Lipoptena cervi*) (Bianchessi et al., 2023; Stuen et al., 2013). In the UK and across mainland Europe, *I. ricinus* remains the pertinent vector (Rikihisa, 2011; Woldehiwet, 2006). *Ap* does however exhibit high levels of genetic diversity, resulting in variations in vector competency, pathogenicity and host tropisms (Atif, 2015; Bown et al., 2009; Jahfari et al., 2014). As a result, different vector species may carry different genetic variants. Jahfari and colleagues (2014) pointed at the circulation of distinct, epidemiologically separate strains of *Ap*, that could be grouped into ecotypes based on host and vector specificity. Four ecotypes of *Ap* can be delineated using fragments of the *groEL* gene with ecotype I suggested as the most common and epidemiologically relevant ecotype for livestock farming (Jahfari et al., 2014; Bianchessi et al., 2023; Gandy et al., 2022). Ecotype II is the second most common ecotype encountered, displaying a preference for infection in populations of roe deer (Jahfari et al., 2014). Ecotype III has rarely been encountered, only found in populations of insectivores and rodents and thought to be primarily vectored by *I. trianguliceps* (Bown et al., 2009). The final and fourth European ecotype has only been detected in birds and as such is associated with the bird tick, *I. frontalis* (Jahfari et al., 2014; Palomar et al., 2015). As mentioned in chapter 1, *groEL* is not the only gene used to fingerprint and delineate *Ap* populations, MLST schemes have also been developed to provide a more robust fingerprint for genetic variants (Huhn et al., 2014). Targets of MLST schemes often include *groEL*, along with a series of additional loci such as *ankA*, *typA*, *recG* etc which provide increased sensitivity to typing methodologies. Huhn and colleagues'

MLST scheme is the most comprehensive in Europe and finds that European isolates of *Ap* generate at least five distinct clades. These clades often agree with those defined by *groEL* alone, identifying a large group akin to ecotype I within the first cluster which includes the European zoonotic strains; a ruminant specialist group akin to ecotype II; a rodent specialist group akin to ecotype III; a bird specialist group akin to ecotype IV; and a fifth small group that specialises in wild ruminants (ecotype V?). Ecotype V has not been sufficiently addressed in follow up papers with many groups opting to work with the notion that four ecotypes are circulating within Europe. To address this problem whole genome phylogenetic trees must be produced that represent the total diversity of *Ap* depicted in Huhn's MLST scheme. This will require not only a large sampling effort, but the development of efficient sequencing methodologies to generate genomes at a low cost from a large number of samples in short period of time. Two distinct methodologies exist for the sequencing of *Ap* strains, tick cell culture and the generation of metagenomic assembled genomes (MAGs) directly from infected blood and tissue. At the time of writing, the capacity to sequence *Ap* directly from blood or tissue has not been fully established (Barbet et al., 2013). Therefore, tick cell culture was chosen to sequence *Ap* strains from the UK.

Tick cell culture remains one of the most effective techniques employed for studying *Ap* across the globe (Silaghi et al., 2011; Massung et al., 2007). Two tick cell lines derived from *Ixodes scapularis*, ISE6 (Munderloh et al., 1999) and IDE8 (Munderloh et al., 1994), have proved useful for *in vitro* cultivation and, as a result, have been extensively used in research (Silaghi et al., 2011; Massung et al., 2007). Cultivation of *Ap* is particularly useful when the goal of an experiment is to generate draft and complete genome sequences for isolates, as concentrated cultures improve the proportion of reads belonging to *Ap* when compared to metagenomic sequencing directly from blood and tissue (Barbet et al., 2013). Despite the significant genetic and epidemiological variation amongst strains of *Ap* and advancements in cultivation techniques, limits in the availability of whole genome data still remain. The establishment of isolates of *Ap* in tick cell culture requires time and expertise and may not be possible for diverse isolates such as those that exist in rodents and birds. Although *groEL* represents a highly effective target for discriminating populations of *Ap*, this data alone is insufficient for characterising epidemiologically separate



ecotypes as different species. Therefore, *Ap* remains a single highly diverse species of blood-borne bacteria. Although *Ap* was discovered in Scottish sheep (Gordon et al., 1932), no whole genome representations of *Ap* have been generated with origins in the UK, representing a significant gap in knowledge. Thus, the aims and objectives of the work carried out in chapter 3 was to generate the first complete representations of *Ap* strains isolated in the UK adding to the growing database of genomes available globally. Utilise a range of contemporary bioinformatical techniques to investigate genomic diversity of *Ap* with a focus on key virulence factors such as the intriguing *msp2* surface antigens and the *virB/D4* T4SS discussed in chapter 1. For example, strains of *Ap* can contain over 100 *msp2* pseudogenes that are expressed in a single expression site enabling the evasion of the hosts immune response (Battilani et al., 2017).

## 3.2: Methods

Seven strains: “Old Sourhope” (OS) (Foster & Cameron, 1970) and “Feral Goat” (FG) (Scott & Horsburgh, 1983), Harris, Perth, ZW122, ZW129 and ZW144 which have been previously isolated in ISE6 and IDE8 cells (Woldehiwet et al., 2002; Woldehiwet and Horrocks, 2005) were sequenced using both long and short read systems. These *Ap* strains have been maintained at the Liverpool Tick Cell Biobank by Dr Leslie Sayki who kindly provided cultures for sequencing.

### Collection & Isolation of Strains

The Harris strain was first isolated from a Scottish sheep in the Outer Hebrides of Harris, Scotland and cryopreserved as a blood stabilate. Isolation was achieved by Dr G.R. Scott and Miss D. Horsburgh in 1982 in collaboration with Mr. A. Whitelaw and the Hill Farming Research Organisation. Harris was noted to be more virulent than most other strains (Centre for Tropical Veterinary Medicine [CTVM], 1983). Harris was then established in ISE6 tick cell lines as described by Woldehiwet et al., 2002 and stored at -114°C until required. The strain was then donated to the Tick Cell Biobank, Liverpool which maintained and proved material for whole genome sequencing.

The Old Sourhope (OS) strain was originally collected from blackface sheep in the Southern Uplands of Scotland in the neighbouring counties of Selkirkshire and Peeblesshire (Foster & Cameron, 1970). Blood stabilates were collected and pooled from blackface ewes purchased from the farms and passaged through susceptible sheep. The infected animals presented with typical TBF symptoms, characterised by fever and neutropenia. Blood stabilates were maintained and occasionally passaged in sheep until the strain was established in IDE8 tick cells (Woldehiwet et al., 1987; Woldehiwet et al., 2002). The strain was then donated to the Tick Cell Biobank, Liverpool which provided the source material for whole genome sequencing.

The Perth strain was first isolated from Blackface sheep on the hamlet of Amulree in Perth, Scotland and cryopreserved as a blood stabilate (CTVM, 1979). The strain was then established in ISE6 tick cell lines as described by Woldehiwet et al., 2002 and stored at -114°C until required. The strain was then donated to the Tick Cell Biobank, Liverpool which maintained and provided material for whole genome sequencing.

The Feral Goat (FG) strain was originally isolated from feral goats on the Galloway hills in the south-west of Scotland and cryopreserved as blood stabilates with routine passages through sheep at the Centre for Tropical Veterinary Medicine in Edinburgh (Scott & Horsburgh, 1983; CTVM, 1982). FG was noted to be significantly more virulent in sheep than in goats (CTVM, 1983). FG was then grown and isolated in continuous IDE8 tick cell lines as previously described (Woldehiwet et al., 2002) and then donated and maintained at the Tick Cell Biobank, Liverpool which provided material for whole genome sequencing.

ZW122, ZW129 and ZW144 were isolated from Welsh sheep, English sheep and English cattle respectively by Professor Zerai Woldehiwet, University of Liverpool. Initial investigations into the strains were carried out by Bown et al., 2007 which identified unique VNTR loci. Professor Zerai Woldehiwet established the strains in tick cell culture where they have remained preserved at the tick cell biobank. Genetic material was provided from these revived cultures for sequencing.

Cell Culture (Dr Leslie Sakyi)

*Anaplasma phagocytophilum* strains were grown at the tick cell biobank, Liverpool in the *Ixodes scapularis* cell line IDE8 (Munderloh et al., 1994) as described previously (Woldehiwet et al., 2002) with some modifications. Briefly, IDE8 cells were grown in sealed, flat-sided culture tubes (Nunc, Thermo-Fisher) at 32°C in L-15B medium (Munderloh and Kurtti, 1989) supplemented with 10% tryptose phosphate broth (Invitrogen), 10% heat-inactivated foetal bovine serum (Invitrogen), 0.1% bovine lipoprotein concentrate (MP Biomedicals, Thermo-Fisher), 2 mM L-glutamine (Sigma), 100 units/ml penicillin and 100 µg/ml streptomycin (Sigma). Medium was changed weekly by removal and replacement of  $\frac{3}{4}$  of the medium volume. Cultures were monitored by weekly inverted microscope examination and preparation of Giemsa-stained cytocentrifuge smears. When the *Ap* infection became heavy (>80% of cells infected, cytopathic effect visible), the bacteria were passaged onto fresh IDE8 cells by transfer of 0.3 ml infected culture suspension.

Semi-purified *A. phagocytophilum* were harvested from infected IDE8 cultures by a modification of a published method (Ferreiro et al., 2016) as follows. The cells and supernatant medium were centrifuged at 1,500 x g for 5 min, the supernatant was saved, the cell pellet was resuspended in 0.5 ml of trypsin solution (500 µg/ml in PBS) and incubated at 37°C for 20 min. The saved supernatant was added to the cell suspension which was then passed ten times through a bent 26-gauge needle to disrupt the cells. The resultant suspension was centrifuged at 1,500 x g for 5 min, the supernatant was collected and centrifuged at 15,000 x g for 5 min to pellet the bacteria. The bacterial pellets were held on ice until used for DNA extraction.

### DNA Extractions

The Promega Wizard HMW DNA extraction kit (A2920) was utilised to extract DNA from tick cell cultures. Cultures were harvested at maximum bacterial load and DNA extractions were performed following manufactures instruction for isolating HMW DNA from tissue culture cells.

### MicrobesNG

Short read sequencing for four *Ap* strains (ZW122, ZW144, Perth, and OS) was outsourced to microbesNG for Illumina 2 x 250bp paired end sequencing on a

NovaSeq platform. For each strain 1ng aliquots of DNA in 100µL of elution buffer were dispatched to microbesNG. The standard short read service was purchased which aimed to provide 30X coverage of samples.

### Illumina MiSeq

Short read sequencing with the Illumina MiSeq was performed in-house with MiSeq v2 reagent kits on just FG as a pilot for further sequencing attempts by Dr Ian Goodhead. *Ap* FG was sequenced using a v2 50-cycle kit (MS-102-2001) following manufacturer's instructions. *Ap* Harris and *Ap* ZW129 were sequenced as part of a larger run with a 300-cycle kit (MS-102-2002) following manufacturer's instructions by myself. Libraries were indexed with the Nextera XT index kit v2 Set A (FC-131-2001) and quantified throughout using the Agilent tapestation 2200 with D1000 high-sensitivity screentape and reagents, and the Qubit 3.0 fluorometer using the dsDNA high sensitivity assay kit. All quantification steps were performed following manufacturer's guides. Data was uploaded and retrieved from Illumina BaseSpace which deposited reads in Fastq format.

### Oxford Nanopore minION

Long reads for all strains of *Ap* were generated in-house using the ONT platform. Genomic DNA libraries were prepared with ligation sequencing kit SQK-LSK109 and barcoded with the native barcoding expansion 1-12 (EXP-NBD104) according to manufacturer's instructions. The barcoded libraries were quantified on the qubit 3.0 fluorometer using the dsDNA high sensitivity assay kit and measured on the Agilent tapestation 2200 using genomic DNA screentape and reagents following the manufactures guides. Quantified libraries were then normalised and pooled accordingly. The pooled library was loaded into an R9.4.1 minION flow cell connected to the minION MK1C system and ran on default settings for 24 hours. The onboard minKNOW 22.08.9 software basecalled reads in real-time using the fast-basecalling method, before depositing read data in fastq format.

### Assembly & Annotation

Short and long read datasets were compiled on the University of Salford Biol-2 server for processing. Long read data were concatenated prior to the removal of adapters using the publicly available PoreChop software (Wick et al., 2019). Reads

were then processed with the Filtlong software package (GitHub: rrWick/FiltLong). A minimum read length of 1000 was set to remove short reads from the long read dataset. Short read data was similarly processed using the short read optimised software package, FastP to both filter and trim reads based on quality and remove sequencing adapters. Short read datasets then underwent filtering based on identity to a draft *Ixodes scapularis* genomes (GCA\_016920785.2). Unaligned reads were exported whilst aligned reads were discarded. Processed short and long read datasets were assembled with the unicycler assembly pipeline which constructs genomes with the short read data set first before bridging fragments with long reads. Assembled genomes were annotated with three different annotation packages: RASTtk, PROKKA and ggCaller.

### Phylogenetics

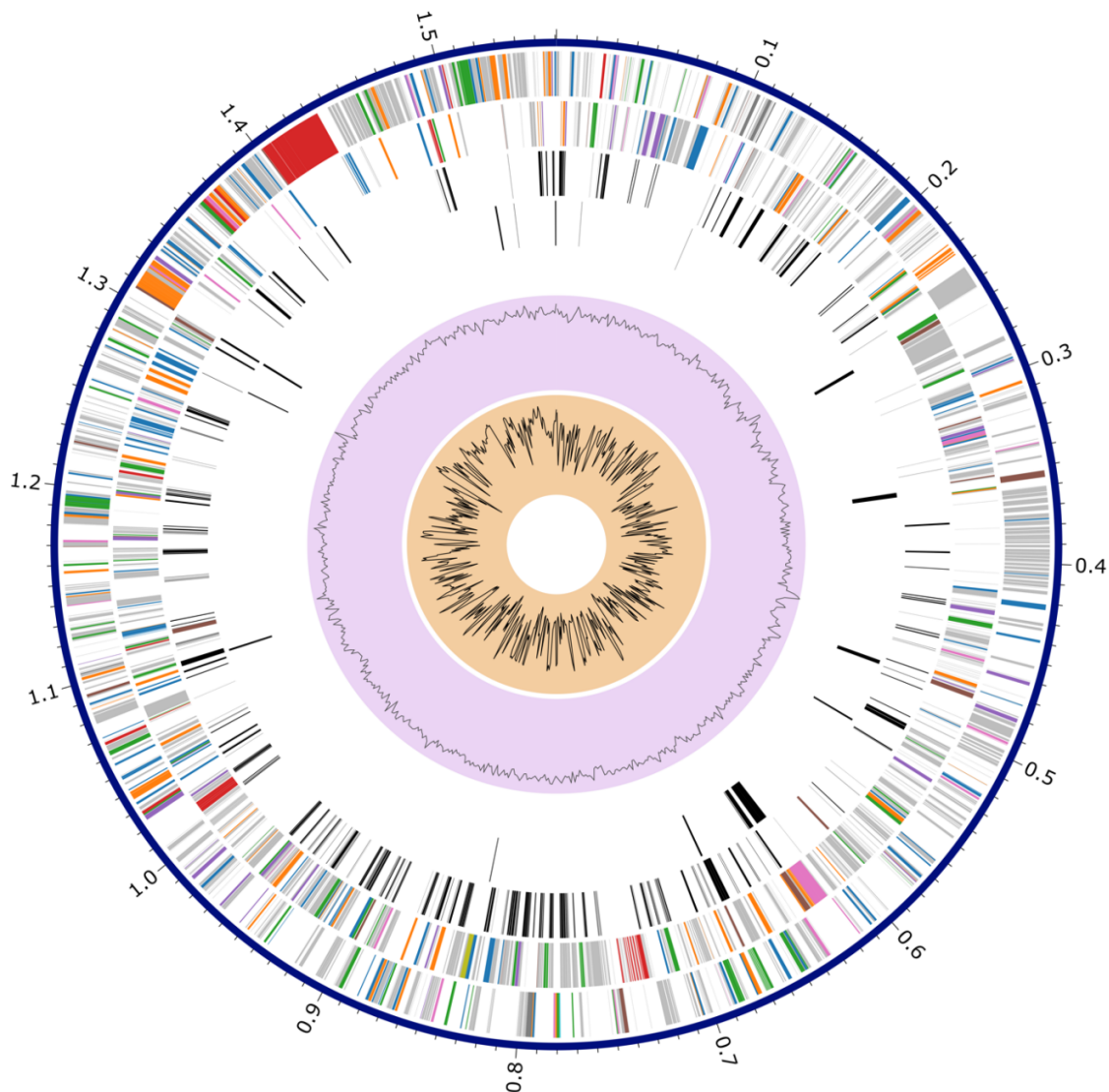
Whole genome phylogenetic trees were constructed based on protein sequences of aligned and concatenated single copy core genes. Annotations generated from RASTtk were used to identify and compile core genes. The Bacterial and Viral Bioinformatics Research Centres (BV-BRC) bacterial genome tree pipeline was employed for generating 500 replicate maximum-likelihood trees. The pipeline takes annotated genomes, identifies single-copy PGFams and processes them with RAxML. Phylogenetic trees were generated in newick format and manually processed into figures using the Geneious Prime software package and Inkscape.

### Pangenome & Average Nucleotide Identity

The *Ap* pangenome was generated using the pangenome pipeline available in the Anvi'o software package. Anvi'o was provided with GenBank files annotated with ggCaller which were parsed into fasta assemblies and external gene calls using native Anvi'o scripts. A local html session was used to evaluate and annotate the pangenome data. PyANI was used to calculate an ANI matrix of all inputted genomes (Pritchard et al., 2016). Results were output as SVGs and visually improved in Inkscape.

### 3.3: Results

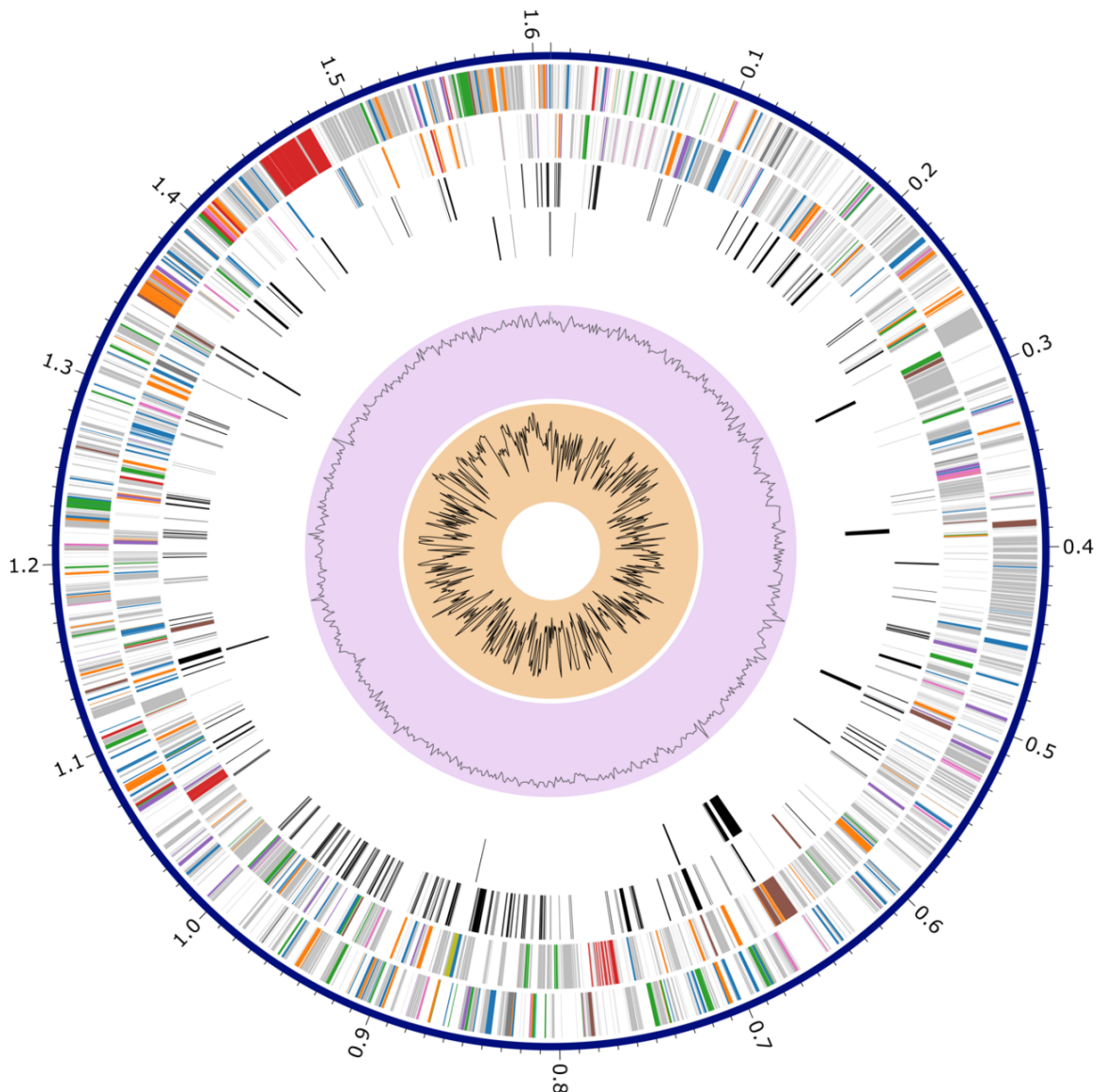
Complete and draft assemblies were produced for all seven strains with varying success. Strains have been ordered in terms of assembly quality (best to worst).



**Figure 24:** A Circos plot of the *Anaplasma phagocytophilum* Harris strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

A complete assembly of the Harris (Figure 24) strain was assembled into a single contig with a length of 1.56Mb and GC content of 41.84%. A total of 1,645 CDS were

annotated across the genome with RASTtk including 267 repeat regions 3 rRNA and 37 tRNA. 704 of the CDS were annotated as hypothetical proteins. The metabolism (blue) and protein processing subsystems accounted for more than half of CDS across the genome and 20 antibiotic resistance genes and 20 transporters making up a complete virB/D4 T4SS were detected.



**Figure 25: A Circos plot of the *Anaplasma phagocytophilum* OS strain.** From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

The OS genome (Figure 25) was assembled into one contig with a length of 1.61Mb and GC content of 41.8%. There were 1357 CDS predicted across the genome 763



of which were annotated as hypothetical proteins. Metabolism and protein processing again accounted for the majority of CDS annotations; 19 antibiotic resistance genes and 17 transporter genes making up a complete *virB/D4* T4SS were detected.

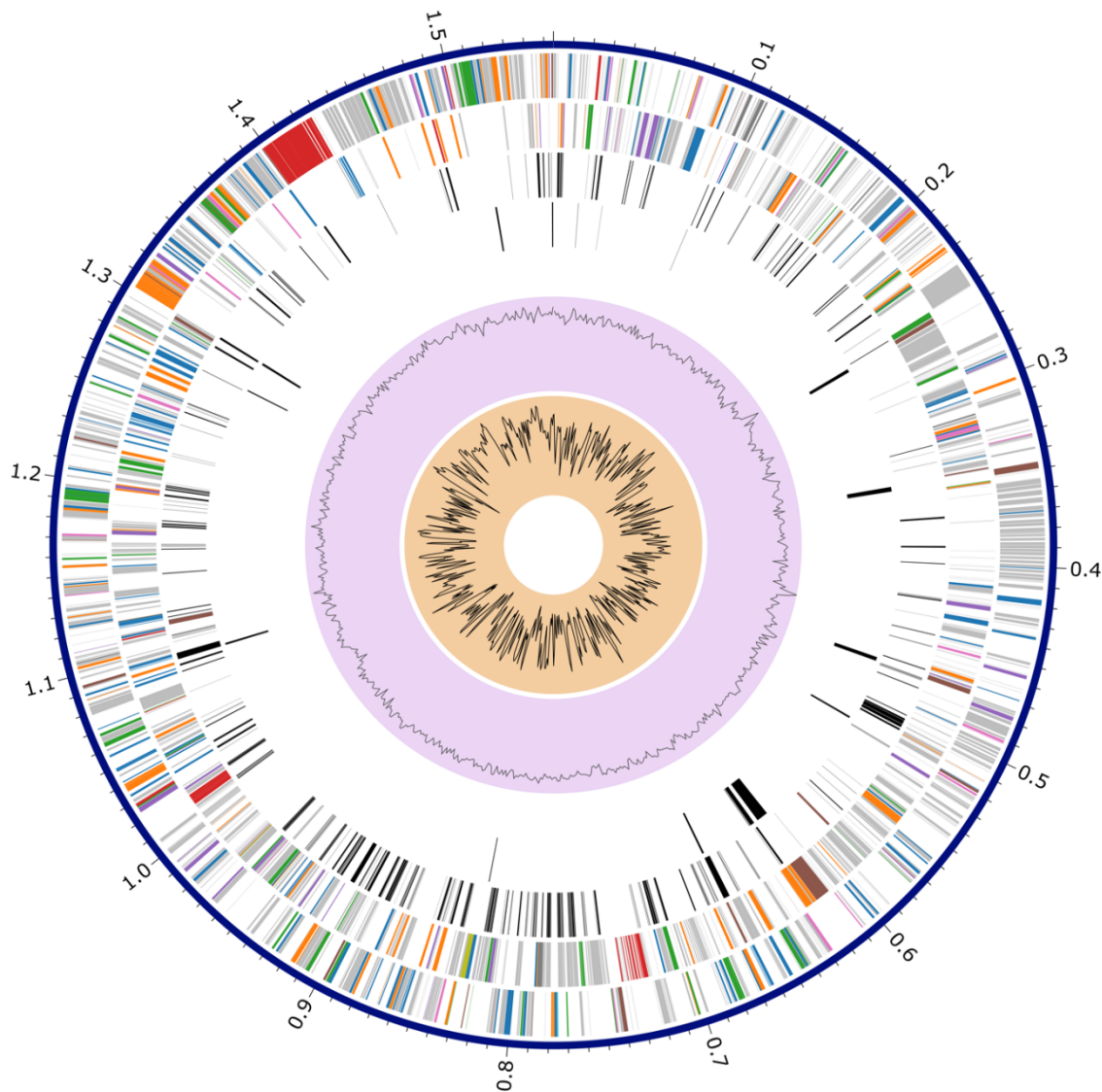
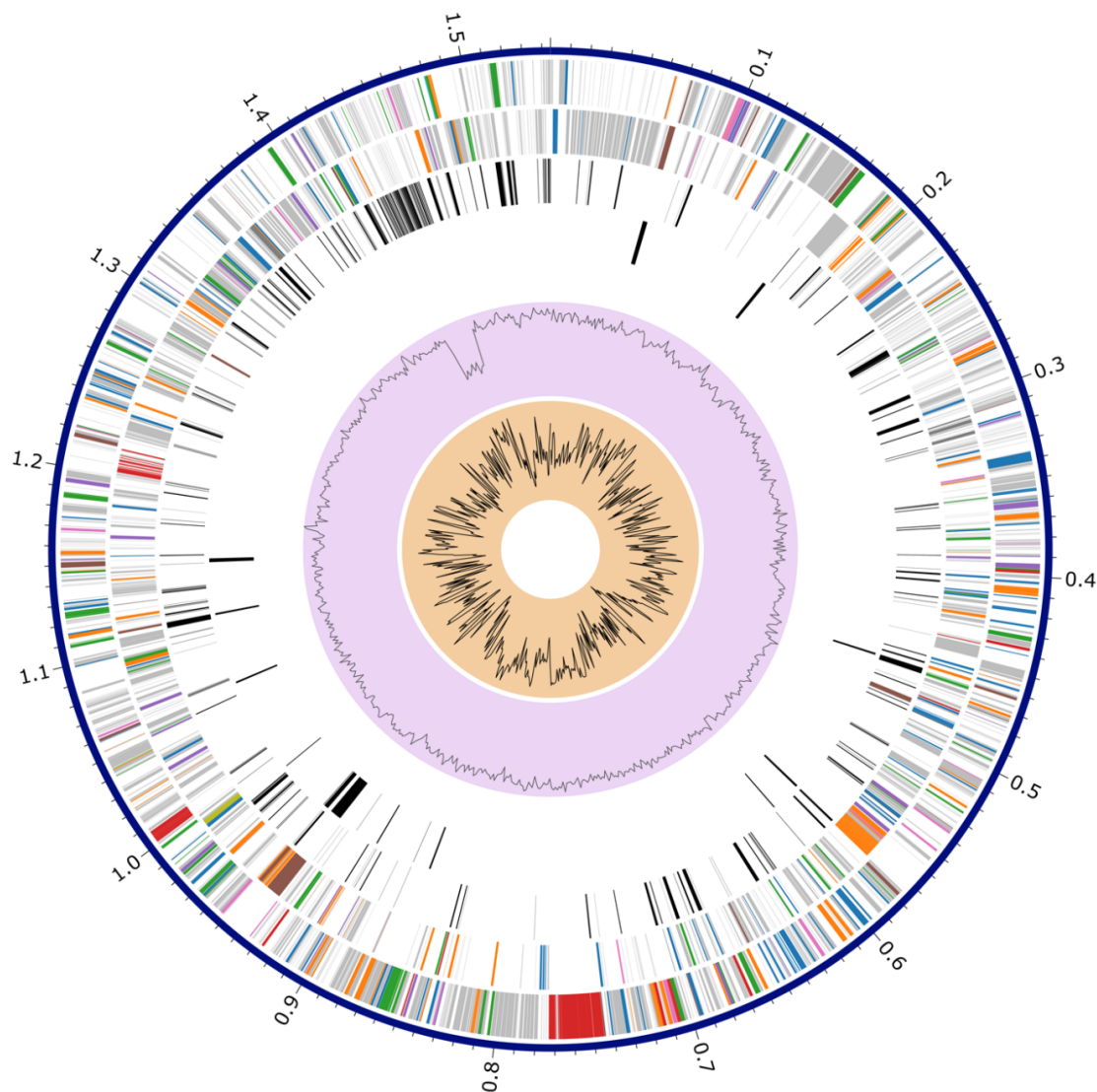


Figure 26: A Circos plot of the *Anaplasma phagocytophilum* ZW129 strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

ZW129 was assembled (Figure 26) into a circular contig measuring 1.55Mb with a GC content of 41.81%. There are a predicted 1632 CDS 703 of which are hypothetical proteins. Two specialty gene types were detected including 15



transporter genes encompassing a complete virB/D4 T4SS and 20 antibiotic resistance genes.



**Figure 27: A Circos plot of the *Anaplasma phagocytophilum* Perth strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).**

The Perth strain (Figure 27) was assembled into 55 contigs spanning a total length of 1.54Mb with a GC content measuring 41.37%. A contig L50 of 3 and N50 of 247kb was achieved. RASTtk predicted 1821 genes, 37 tRNA and 3 rRNA. 906 of the predicted CDS were hypothetical proteins. 19 antibiotic resistance genes and 20 transporter genes making up a complete virB/D4 T4SS were detected.

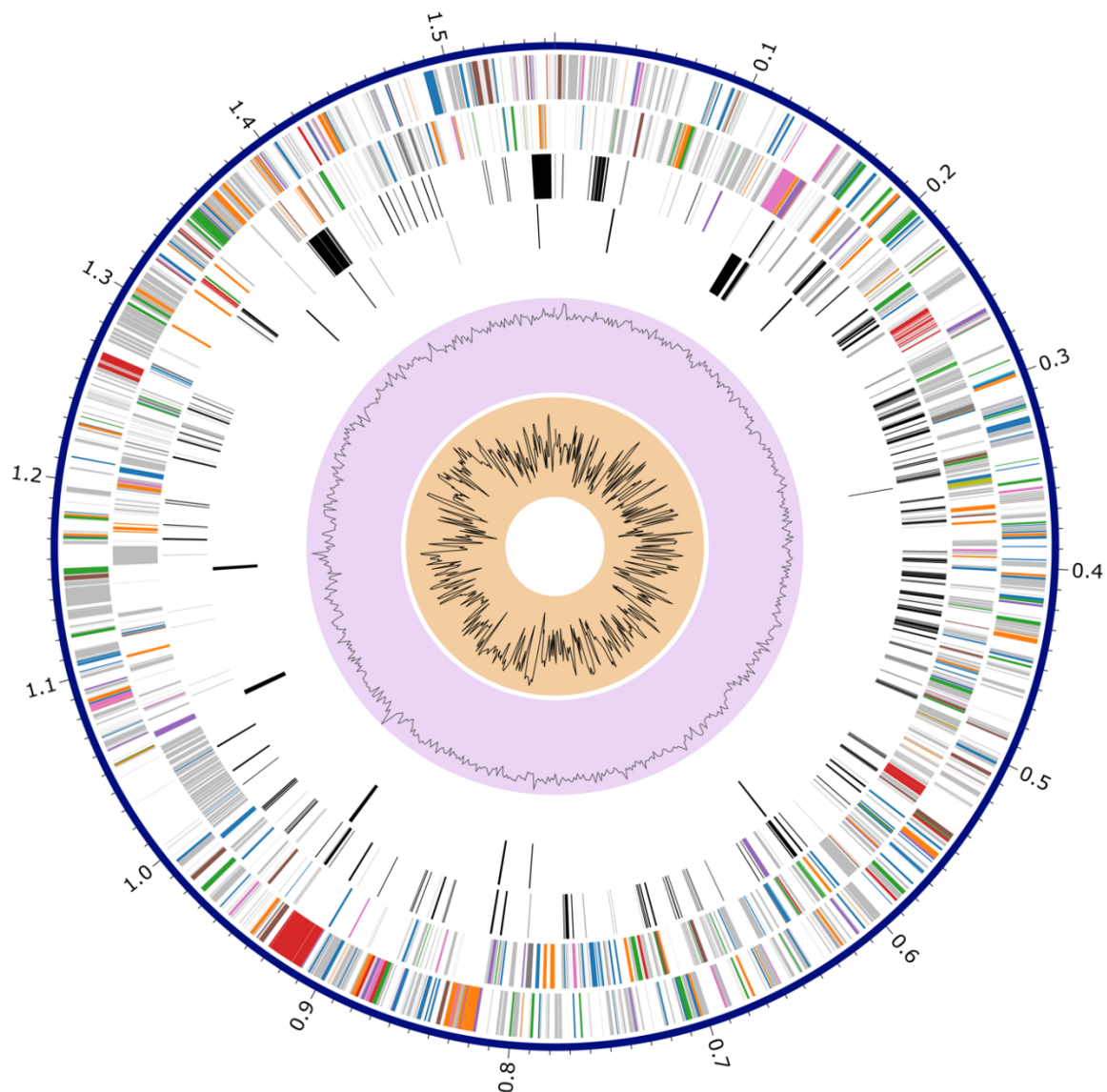
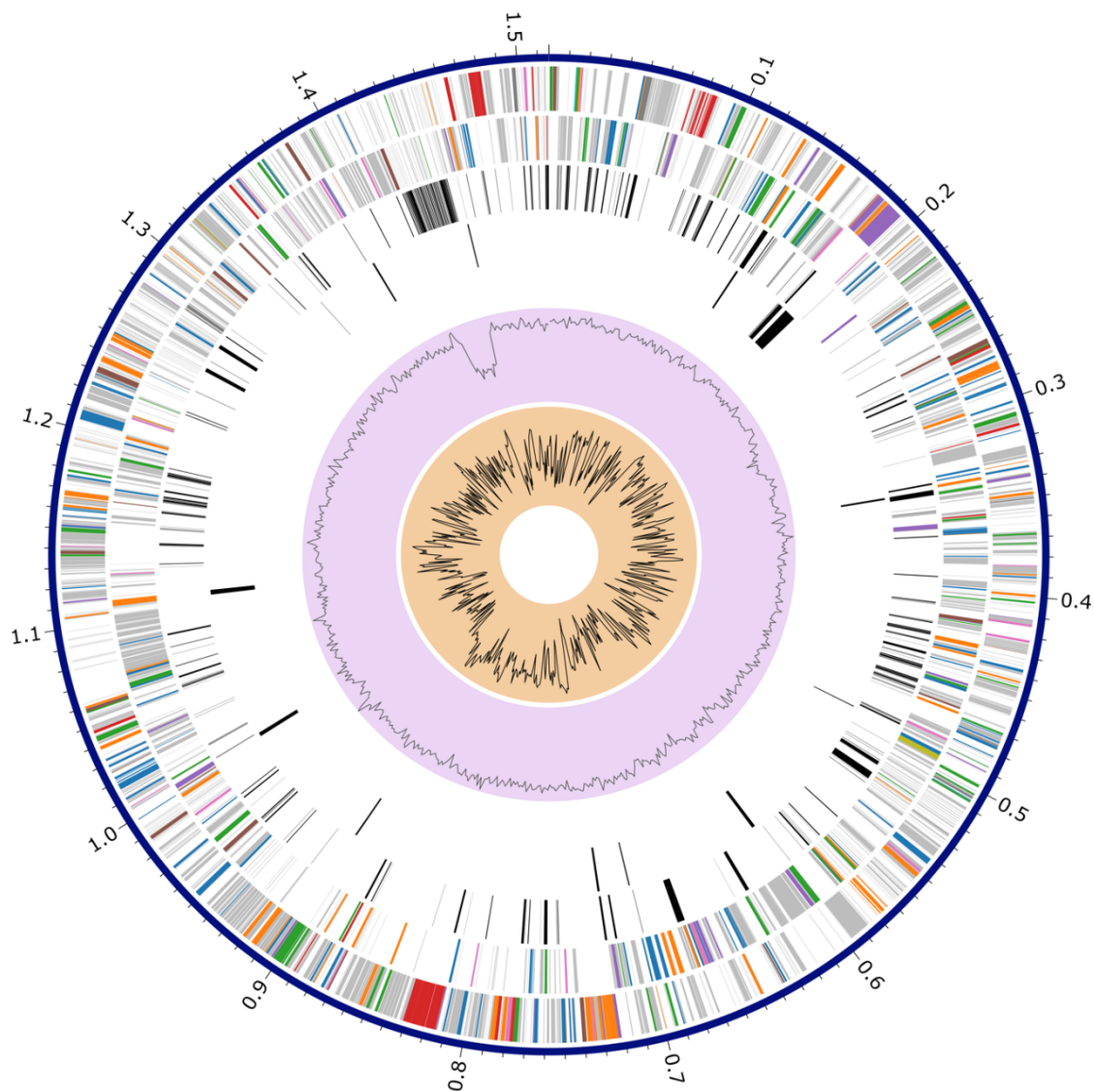


Figure 28: A Circos plot of the *Anaplasma phagocytophilum* ZW122 strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

ZW122 (Figure 28) was assembled into 6 contigs spanning 1.55Mb. A contig L50 of 1 and N50 of 931kb was achieved. A total of 1700 CDS, 38 tRNA and 3 rRNA were predicted. Of the 1700 CDS, 753 had no functional assignment and as such were labelled as hypothetical proteins. 22 antibiotic resistance genes and 19 transporter genes making up a complete *virB/D4* T4SS were detected.



**Figure 29: A Circos plot of the *Anaplasma phagocytophilum* ZW144 strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).**

The ZW144 genome (Figure 29) measures 1.52Mb across 38 contigs with an L50 of 4 and N50 of 154kb. GC content measures 41.29% and RASTtk identified 1776 CDS, 38 tRNA and 3 rRNA. The majority of proteins were given functional annotations (921); 855 CDS were hypothetical proteins. Protein processing and metabolism were the dominant subsystems accounting for just over a quarter of gene functions each. A total of 18 antibiotic resistance genes and 19 transporter genes were identified in the assembly, with a likely functional virB/D4 T4SS.



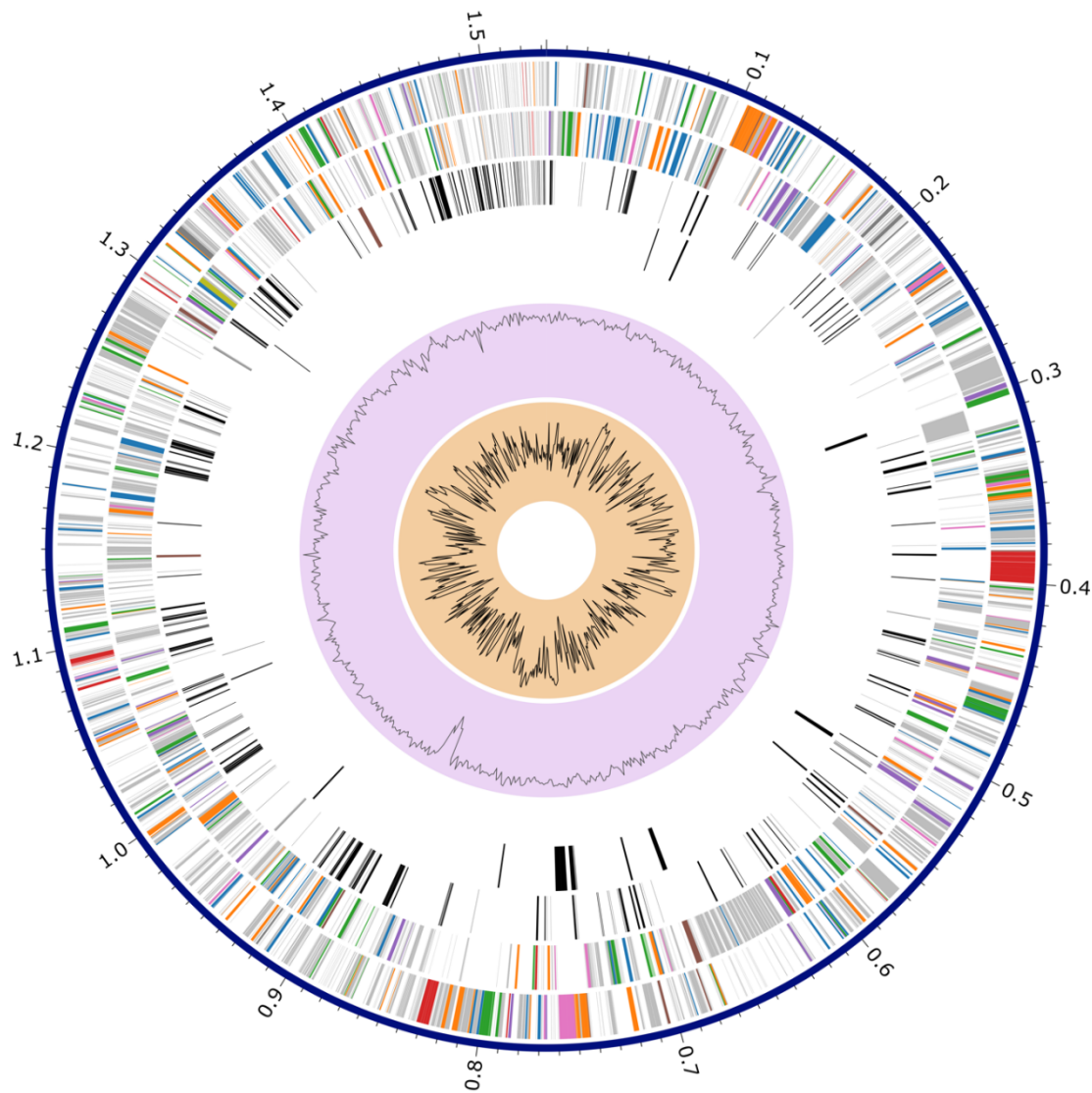
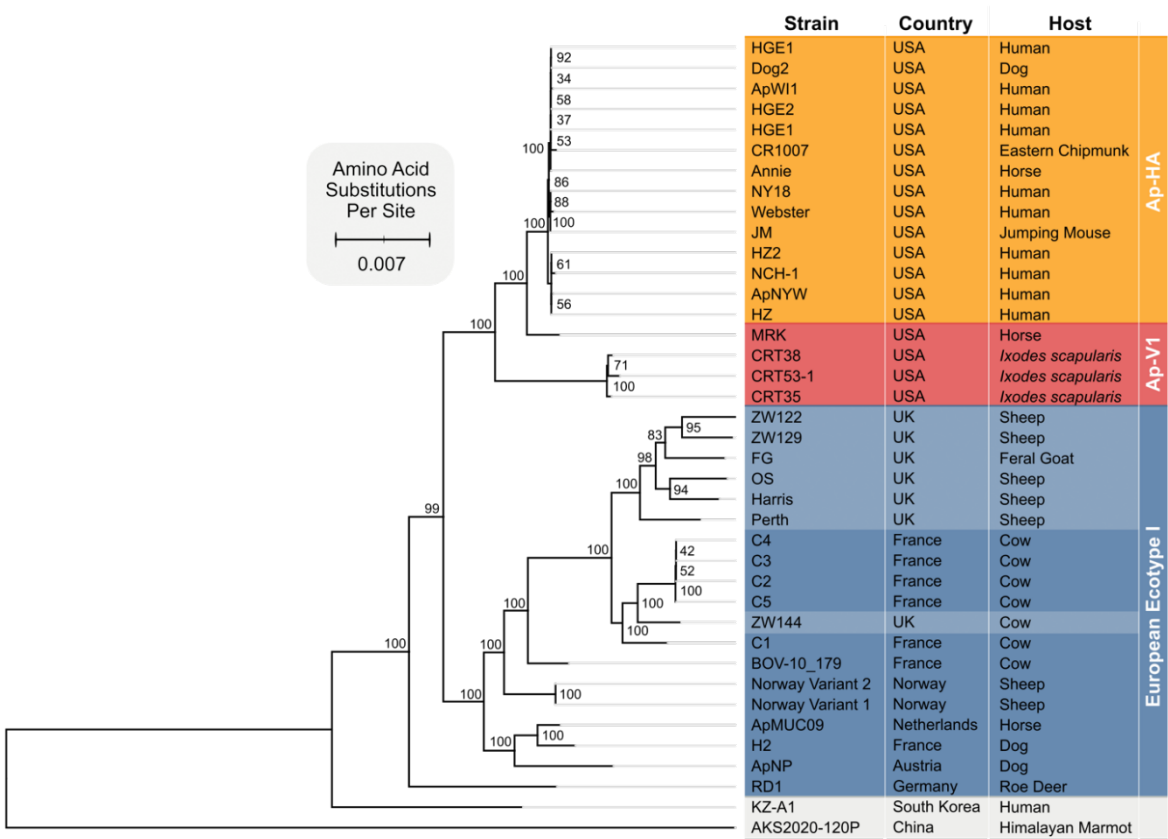


Figure 30: A Circos plot of the *Anaplasma phagocytophilum* FG strain. From outer to inner rings, the plot shows contigs, CDS on the forward strand, CDS on the reverse strand, RNA genes, CDS with homology to known antimicrobial resistance genes, CDS with homology to known virulence factors, GC content, and GC skew. Colours represent the subsystem the genes belong to: (Orange, protein processing); (Blue, metabolism); (Green, energy); (Brown, stress response, defence, virulence); (Pink, RNA processing); (Purple, DNA processing); (Red, membrane transport); (Grey cellular processes); (Yellow, cell envelope); (Light blue, miscellaneous).

The genome of FG (Figure 30) was typical for European ecotype I with GC content measuring 41.58% across 1.53Mb. The assembly for FG is split into 275 contigs with an L50 of 10 and N50 of 61kb. RASTtk predicted 1989 CDS, 36 tRNA and 2 rRNA, of the predicted proteins 1005 were given functional annotations, 984 were hypothetical. Subsystem analysis indicates that protein processing and metabolism represent over half of all gene functions. Some specialty genes were detected

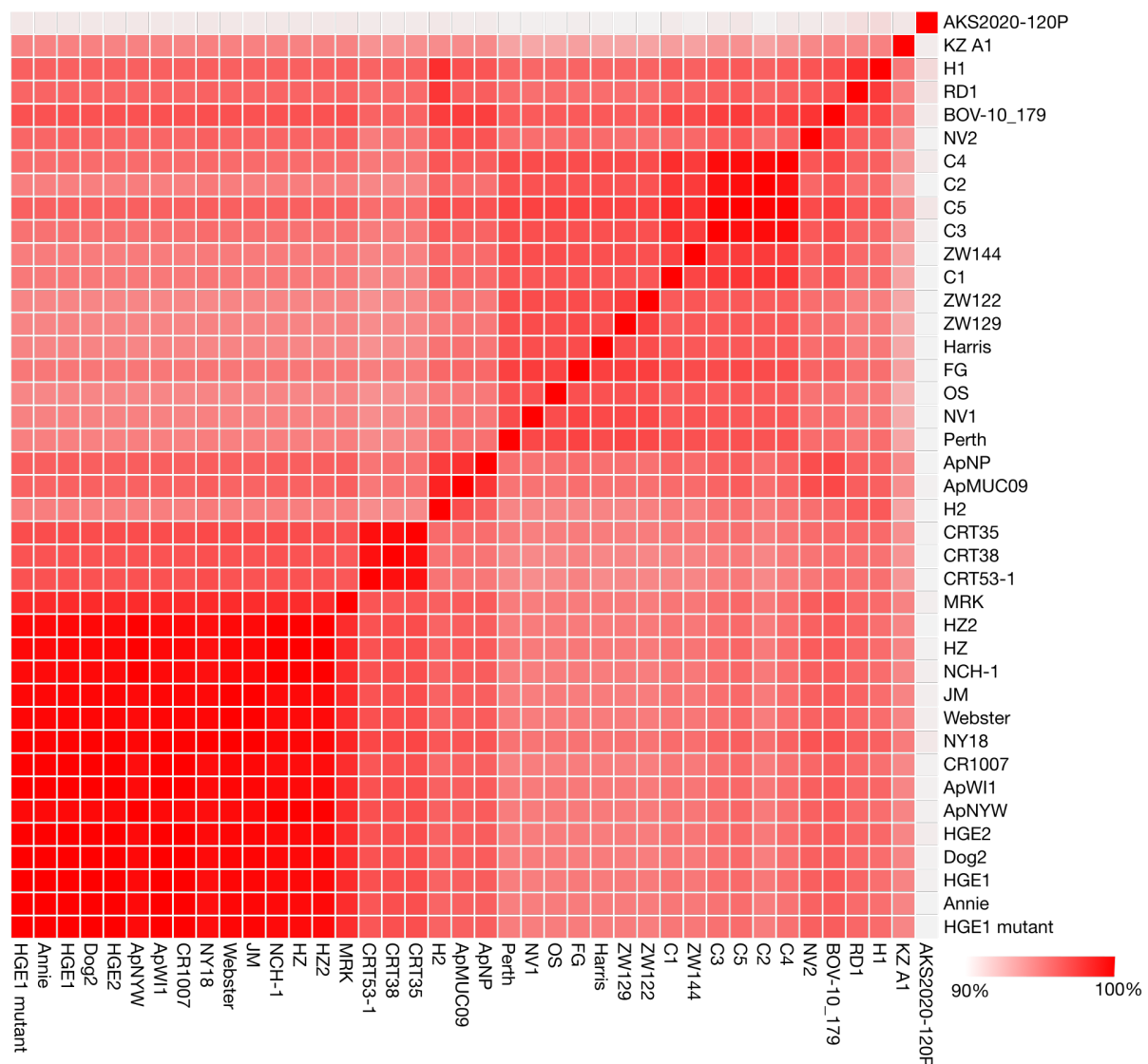
including 19 antibiotic resistance genes and 23 transporters, including a likely functional *virB/D4* T4SS.



**Figure 31: Maximum-likelihood phylogenetic tree generated from a concatenated alignment of amino acid sequences of 500 single copy core genes within the *Anaplasma phagocytophilum* (*Ap*) species, generated with RAxML using the DUMMY2 protein model. Scale bar represents 0.007 amino acid substitutions per site and colour delineate known clusters/ variants. Dark blue represents strains of European origin, determined to be ecotype I. Light blue represents UK derived ecotype I strains, yellow represents the North American human active variants of *Ap* (*Ap*-HA) and red represents the North American variant 1 strains (*Ap*-V1).**

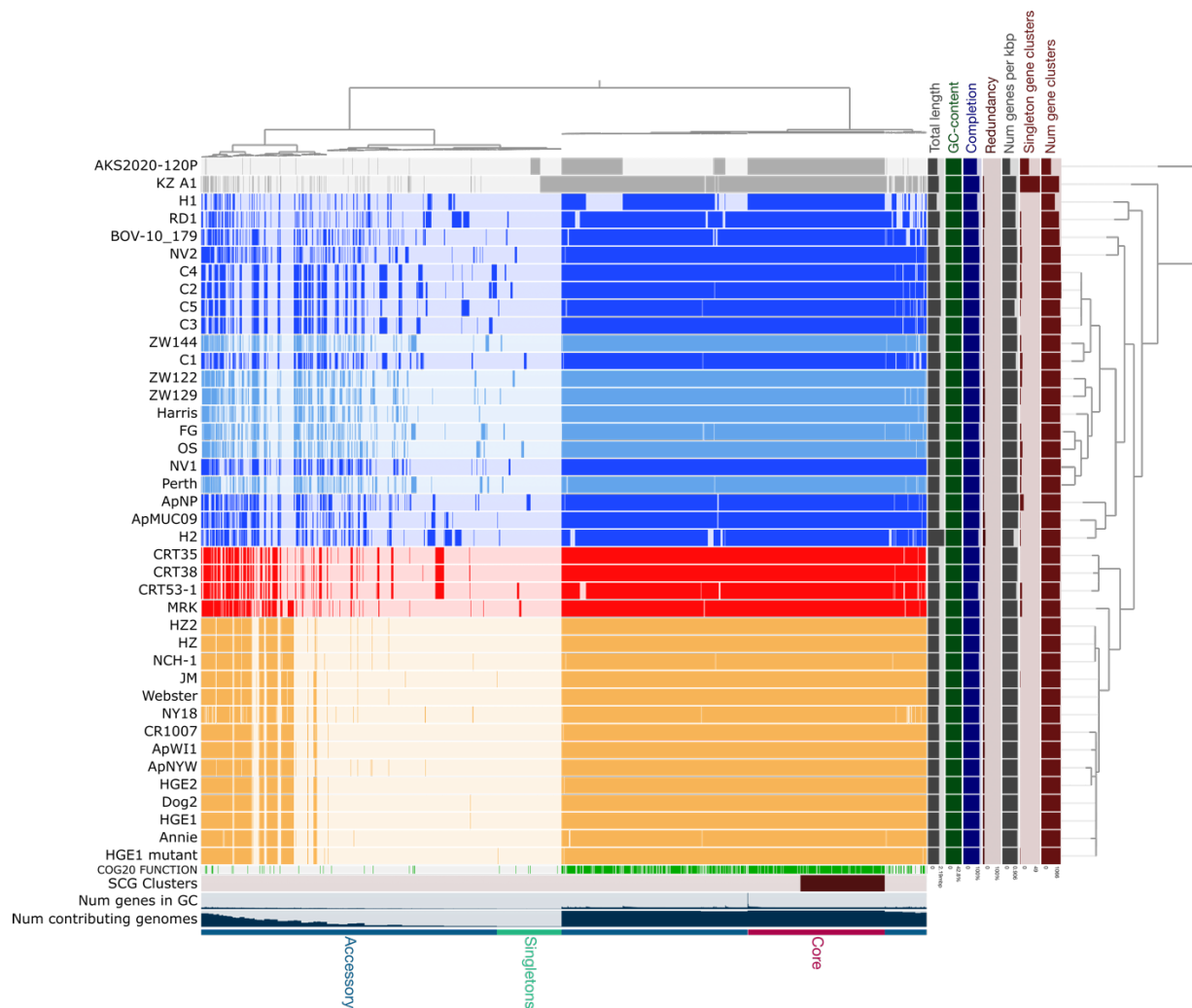
A maximum likelihood tree was generated from the concatenated alignment of 500 single-copy core genes shared across the 40 available *Ap* whole genomes, including seven novel UK-derived genomes. The tree clearly delineated Asian, European, and North American strains and identified a subdivision within North America that separated the largely clonal human-active variants and *Ap*-V1. In Europe, the genetic diversity of *Ap* is notably higher with multiple distinct subclusters, including ruminant-derived UK strains, cattle-derived strains, and dog and horse-derived strains. Interestingly, the Norway sheep variants did not cluster with the UK-derived sheep strains but formed an outgroup on the larger Anglo-French cluster, indicating a geographical influence on *Ap* diversity. The RD1, KZ-A1, and AKS2020-120P

strains represent the most diverse *Ap* genomes with significantly higher genetic distances from the main European and North American clusters.



**Figure 32: Average nucleotide identity heatmap of 40 *Anaplasma phagocytophilum* genomes calculated with pyANI.**

ANI reveals high genetic relatedness across most *Ap* genomes, including those identified as European Ecotype I and both North American ecotypes. However, novel species classification boundaries (<95% identity) are crossed when comparing the Himalayan marmot strain AKS2020-120P or the KZ-A1 strain with the well-represented western variants or with each other, as ANI drops to ~90% and ~93%, respectively.



**Figure 33: Pangenome analysis of *Anaplasma phagocytophilum* generated with Anvi'o displaying 40 strains. The layers represent individual genomes organised by phylogenetic relationships based on 500 single copy core genes. Colours represent distinct *Ap* clusters. Genome completeness, redundancy, GC content, length, genes per kbp singletons, and total number of gene clusters are indicated on the right. On the bottom NCBI COG20 functions, number of genomes that contain a specific gene cluster and number of contributing genomes is indicated. In addition to this, the core, accessory and singleton gene clusters are highlighted with red, blue and green respectively.**

The pangenome analysis identified 1,660 homologous gene clusters containing 9,910 total genes across the examined *Ap* strains. A significant proportion, 77.49%, of these genes were annotated as hypothetical proteins with no assigned functional annotations. Out of the 20 amino acids *Ap* appears to only encode for enzymes responsible for proline, glutamine, glycine and aspartate biosynthesis. The glycolysis enzymes present are reduced to those that produce glyceraldehyde-3-phosphate and dihydroxyacetone phosphate implying *Ap* requires glycolytic intermediates from the host/ vector. In addition to this several proteins exist such as those that make up multidrug efflux pumps that may contribute to general stress responses, defence against host derived compounds and antibiotics or may even facilitate horizontal

gene transfer. Additional proteins are present that contribute to crucial functions such as DNA repair, replication and recombination such as DNA translocases. These include *recA*, *ftsK*, *ruvAB* and the *virB/D4* T4SS which likely functions to maintain and evade infection in neutrophils through the translocation of effector proteins. Of these proteins *ankA*, *ats-1*, *ampA*, *hapE* and *ompA* were detected across the species. Five distinct groups were observed within the pangenome: European ecotype I, *Ap*-HA (North American), *Ap*-V1 (North American), KZ-A1, and AKS2020-120P strains as previously described in phylogenetic analyses. The European ecotype I strains displayed higher genomic diversity compared to the North American *Ap*-HA and *Ap*-V1 variants, which appeared largely clonal within their respective groups based on gene content similarity. While the *Ap*-V1 strains were distinct from the *Ap*-HA strains and the European ecotype I strains, all North American and European strains shared a significant core and soft-core genome. Gaps were present in the flanking accessory genome regions of the core genome (soft-core) in H1 and AKS2020-120P due to poor assembly quality and the strict annotation methodology employed by ggCaller. The European variants possessed larger accessory and singleton genomes compared to the North American strains. Both divergent strains KZ-A1 and AKS2020-120p shared the majority of their gene content with other *Ap* strains however they possess larger singleton genomes. GC content remained stable across all strains fluctuating between 41.2% and 42.9% with AKS2020-120P exhibiting the lowest and C5 the highest. Corrected anvio-7.1 genome completeness varied from 84.65% - 98.73% whilst redundancy ranged from 0% - 14.08%. The majority of assemblies were good quality as defined by high completion and low redundancy. H1, and AKS2020-120P were determined to be poor quality. The number of gene clusters and subsequently gene per kb was highly dependent on assembly quality with the poor-quality assemblies and those split into many contigs suffered from reduced annotation of CDS using the ggCaller annotation algorithm.

---

### 3.4: Discussion

The genetic and epidemiological diversity of *Ap* continues to pose a significant challenge to both animal health and agricultural productivity (Brodie et al., 1986; Woldehiwet, 2006), particularly in areas of high tick prevalence in the UK (Lihou et



al., 2020). Prevalence of ticks and *Ap* infections within ticks may not provide ample information for quantifying veterinary or zoonotic risk as detailed information on strain identity is required. In addition to this, the presence or absence of wildlife species susceptible to certain strain of *Ap* can also drastically impact infection risk amongst other factors such as farm location, temperature and exposure to the sun which can all impact tick density. This makes quantifying risk difficult without a highly sensitive and efficient fingerprinting process to identify *Ap* at the strain level and a comprehensive screening methodology for domesticated livestock, wildlife and ticks. This study aimed to contribute to the understanding of *Ap* strains circulating within UK livestock by generating the first complete genomic representations of seven isolates using a combination of long and short read sequencing technologies. By focusing on seven strains previously isolated in tick cell cultures, this study was able to generate three complete genome sequences (Harris, OS and ZW129) and four draft genome sequences (Perth, ZW122, ZW144 and FG) of varying sequence fidelity. Our findings not only highlight the genetic diversity present within UK *Ap* strains circulating in livestock but also emphasise the critical need for more comprehensive genetic typing approaches beyond the use of individual markers such as *groEL* or MLST schemes incorporating a limited number of markers in order to more comprehensively delineate the ecotypes of *Ap* present in the literature. Through whole genome analysis of the UK livestock strains and all available whole genome data for the species, I was able to gain insights into the metabolism and virulence factors driving *Ap* evolution. These include the abundant *msh2* pseudogenes, *virB* T4SSs and its diverse effector proteins that contribute to immune evasion, modulation and varying pathogenicity across *Ap* strains.

The assembled genomes ranged in size from 1.52 Mb to 1.61 Mb, with GC content varying from 41.29% to 41.84%, which is consistent with the reported high GC content of *Ap* compared to other obligate intracellular organisms in the Rickettsiales order (Battilani et al., 2017). The number of predicted coding sequences (CDS) varied from 1,357 (OS strain) to 1,989 (FG strain). The variation amongst strains is likely down to sequencing quality rather than true content variation with lower quality assemblies possessing an increased number of CDS. When only considering the complete genome assemblies, Harris, OS and ZW129, CDS counts only varied by a maximum of 288 indicating high conservation of genomic features across the

species. Further analysis of this using a pangenome pipeline (Figure 33) revealed that there was low conservation of genome size (1.17Mb – 2.19Mb) which does reduce substantially when removing the two outliers H1 and H2 (1.26Mb – 1.72Mb) generated from French horses (GCA\_900088675.2 and GCA\_900088655.1). Despite variations in genome size, over half of the gene clusters are conserved across most strains indicating high specialisation and streamlining of the genome. There are a number of reasons why bacteria may streamline or oppositely expand their genomes size or contents. For example, genome streamlining can be associated with bacteria that undergo significant environmental stress, no longer need to maintain certain genes for survival, or necessitate the avoidance of host immunity such as intracellular parasites or bacterial symbionts that are increasingly driven by host dependency (Diop et al., 2018; McCutcheon et al., 2012). Genome expansion on the other hand can be associated with bacteria that are self-sufficient, environmentally resistant, and are likely to benefit from rapid adaptation and niche expansion such as those found in soil, water or the gut (Makarova et al., 2001; Diczko et al., 2019). There are of course exceptions to this such as *Legionella pneumophila* an intracellular bacterial parasite that also employs a transmissive environmental phase. Therefore, based on our previously described logic *L. pneumophila* may benefit from both genomic streamlining and expansion. The reality of *L. pneumophila* is that the complex life cycle and requirements for environmental resistance necessitate an expansion of the genome, however, the species has also acquired a number of virulence factors that do contribute to host immune evasion and modulation indicating genome streamlining may not always be the favoured evolutionary trajectory (Joseph et al., 2016). Although no direct evidence of genome streamlining can be observed in European *Ap* strains, the presence of only proline, glutamine, glycine and aspartate biosynthesis enzymes does indeed point to significant host and vector reliance. *Ap* also undergoes significant selective pressures during its obligate intracellular lifestyle that primarily revolve around avoiding detection by host immune cells (Rikihisa, 2011). It is therefore reasonable to assume that strains of *Ap* are constantly undergoing genome streamlining much like their close *Rickettsia* sp. relatives which also boast genome sizes in the range of 1.1Mb – 1.5Mb with coding capacities upwards of 69% (Diop et al., 2018). Interestingly, genomic reductive evolution in *Rickettsia* sp. has been directly associated with an emergence of pathogenicity (Diop et al., 2018; Darby et al.,

2007). Therefore, the observed increase in pathogenicity of *Ap* strains in the USA may in some capacity be linked to the notable reduction in genome size across *Ap*-ha isolates (1.48Mb on average) when compared to their European counterparts (1.60Mb on average). This is even more intriguing when considering the evidence of a recent introduction of European ecotype I strains to North America (Aardema, 2023). The *Ap*-ha haplotype is thought to have directly evolved from these introduced European strains under the influence of a strong genetic bottleneck (founder event) (Aardema, 2023). If true, this would be direct evidence of genome reduction in the species, and a probable link to increased zoonotic pathogenicity.

It would therefore be interesting to explore this in other strains of *Ap* that have significantly reduced genome sizes such as the Chinese strain AKS2020-120P isolated from the Himalayan marmot (1.26Mb) and the currently uncharacterised European ecotypes (II-IV) if variation exists.

Comparative genomics of *Ap* genomes also provides clues into the strategies employed by *Ap*, from host cell invasion to immune evasion and modulation. The *virB/D4* T4SS, a well described transmembrane structure present in all Rickettsiales order bacteria including *Ap* (Crosby et al., 2020; Nui et al., 2010) likely aids in the evasion of host immunity through the translocation of effector proteins. There are several candidate genes that may function as effector proteins including *ankA*, *ats-1*, *ampA*, *hapE* and *ompA* which could be detected in the UK livestock strains. *AnkA* and *ats-1* are abundantly expressed during infection but are not toxic to host cells (Rikihisa, 2011). They contain eukaryotic protein motifs or organelle localisation signals that specific host cell molecules, promoting infection (Rikihisa, 2017). *Ats-1* subverts two important innate immune mechanisms against intracellular infection: cellular apoptosis and autophagy, where autophagy is manipulated to gain nutrients from host cells (Nui et al., 2010; Li et al., 2022). The structure of the *virB/D4* T4SS in *Ap* is homologous to the system encoded by the pTi plasmid in *A. tumefaciens* (Christie, 2004). The distribution and duplication of the *virB/D4* genes within the UK livestock strains is however curious. The components of the system are distributed in three distal clusters as previously described in other strains of the species (Gillespie et al., 2010; Al-Khedery et al., 2012). The first cluster contains *virB8-1*, *virB9-1*, *virB10*, *virB11*, and *virD4*; the second cluster contains *virB2* and *virB4-2*; the third

cluster comprises *virB3*, *virB4-1* and four *virB6* paralogs. *VirB1* is not encoded in *Ap*, a gene encoding a murein-degrading transglycosylase required for channel assembly across the cell wall (Bhatty et al., 2013; Hopper et al., 2005) and *virB5*, which encodes a pilus-associated protein (Aly et al., 2007). *VirB6* and *virB2* have undergone duplication events in *Ap* genomes, for example the Harris genome contains 12 copies of *virB2* with no apparent biological advantage. There are however consistently four copies of *virB6* in *Ap*. *VirB6* is known to confer stability of the T4SS (Kumari et al., 2019), however little is known about the function of expressing four *virB6* paralogs in terms of infection, replication and pathogenesis.

Phylogenetic analysis based on the concatenated alignment of 500 single-copy core genes revealed a clear delineation of Asian, European, and North American strains of *Ap*. Within Europe, the genetic diversity of *Ap* was notably higher, where multiple distinct subclusters within the accepted European ecotype 1 group were observed, including ruminant-derived UK strains, cattle-derived strains, and dog and horse-derived strains (proposed zoonotic group, see Figure 6). Interestingly, the Norway sheep variants formed an outgroup to the larger Anglo-French cluster, suggesting a geographical influence on *Ap* diversity. Variable evolutionary pressures, vectors and hosts across the globe have major impacts on *Ap* evolution (Jaarsma et al., 2019). Geographical location is therefore a key factor that drives the trajectories of *Ap* genome structure. An example of this is the evolution of ecotype II thought to have evolved from the host generalist ecotype I at the end of the last glacial maximum as Europe was recolonised by mammal populations such as roe deer (Aardema et al., 2022). The findings of my phylogenetic tree are for the most part consistent with the previous observations of genetic diversity within *Ap* strains based on *groEL* and MLST typing schemes however, they reveal further delineations within the predominant ecotype I group (Jahfari et al., 2014; Battilani et al., 2017; Huhn). I was unable to assess the existence of multiple European ecotypes using this methodology due to the fact that all strain sequenced in Europe were ecotype I. Therefore, there is an acute need for more whole genome data from diverse *Ap* isolates if the differences in predictions of various typing schemes are to be resolved, necessitating efficient sequencing methodologies.

The average nucleotide identity (ANI) analysis revealed high genetic relatedness across most *Ap* genomes, including those identified as European Ecotype I and both North American ecotypes. However, novel species classification boundaries were crossed when comparing the Himalayan marmot strain AKS2020-120P and the KZ-A1 strain with the well-represented western variants or with each other, suggesting the potential existence of distinct species or subspecies within the *Ap* complex. Crossing arbitrary ANI values alone is not sufficient evidence for species separation but other factors such as ecology, phenotype, reproductive isolation, gene flow and phylogenetic structure needs to be combined in a holistic approach. The specialisations of ecotypes II, III and IV already described provide additional weight to the multiple species' hypothesis, however I believe it is important to comprehensively address the grey areas between these ecotypes, particularly ecotype I and II prior to any reclassifications. The pangenome analysis identified 1,660 homologous gene clusters containing 9,910 total genes across the examined *Ap* strains. A significant proportion (77.49%) of these genes were annotated as hypothetical proteins, emphasizing the need for functional characterization of the *Ap* genome. The European ecotype I strains displayed higher genomic diversity compared to the North American *Ap*-HA and *Ap*-V1 variants, which appeared largely clonal within their respective groups based on gene content similarity. This finding aligns with the phylogenetic analysis and further supports the notion of higher genetic diversity within European *Ap* strains.

The observation of a geographical influence on *Ap* diversity, as evidenced by the separation of Norway sheep variants from the Anglo-French cluster, suggests the potential for region-specific patterns of *Ap* evolution and host adaptation. Further sampling and genomic characterisation of *Ap* strains from diverse geographical regions could provide valuable insights into the ecological and evolutionary drivers shaping *Ap* diversity. While the present work identified the presence of key virulence factors, such as the *msh2* surface antigens and the *virB/D4* T4SS, a more in-depth analysis of these factors could contribute to our understanding of *Ap* pathogenesis and host-pathogen interactions. Comparative studies of virulence factor diversity across *Ap* strains, coupled with functional characterisation, could provide valuable insights into strain-specific pathogenic mechanisms and potential targets for therapeutic interventions. Integrating the genomic data generated in this study with

epidemiological and clinical data from *Ap* infections in livestock and humans could help elucidate potential links between genetic diversity, host tropisms, pathogenicity, and disease outcomes. Such integrated analyses could inform risk assessment, disease management strategies, and the development of targeted diagnostic and therapeutic approaches.

In conclusion, the present work has significantly contributed to the understanding of *Ap* genetic diversity in the UK by generating the first complete genome representations of seven *Ap* strains isolated from ruminants and ticks. The findings highlight the high genetic diversity of *Ap* within Europe, the potential existence of distinct species or subspecies, and the need for further functional characterisation of uncharacterised genes. Future research should focus on addressing the highlighted areas for expansion, including functional characterisation, investigation of divergent strains, exploration of geographical patterns, characterisation of virulence factors and host-pathogen interactions, and integration with epidemiological and clinical data. By addressing these areas, our understanding of *Ap* biology, epidemiology, and host-pathogen interactions can be significantly enhanced.

# CHAPTER 4

## 4.0: Development & Optimisation of Enrichment Protocols for High-Resolution Sequencing of *Anaplasma phagocytophilum* Genomes from Blood & Tissue.

---

### 4.1: Introduction

Genomic data provides critical information on bacterial diversity, evolution and epidemiology, which can be used to guide public and veterinary health policy (Brand et al., 2008). Whole genome datasets are particularly useful as they allow researchers to determine transmission dynamics at various scales, for example, the emergence of particular strains or pathogenic traits (Koa et al., 2014; Dennis et al., 2022). Whole genome sequencing is therefore routinely implemented for the characterisation of a wide range of bacteria. These methodologies are typically applied to readily culturable bacteria as achieving a sufficient quantity and quality of DNA is relatively simple. A reliance on a culture first step does however represent a major hurdle when working with limited equipment or difficult-to-culture or unculturable bacteria. When a culture-first step is omitted, the result is a metagenomic sequencing experiment that directly sequences DNA from infected material whether that be environmental, or host derived.

Metagenomic whole genome sequencing of microbial organisms infecting host or vector tissue is an exciting area for discovery with far reaching implications for clinical diagnostics, disease management and epidemiology. A major barrier to this approach is the often-overwhelming ratios of host to pathogen DNA in samples with low pathogen abundance (Thoendel et al., 2016; de Albuquerque et al., 2022). For example, only 0.00046% of reads (475 out of 10,196,620) came from *Leptospira* species when whole genome sequencing was used to diagnose neuroleptospirosis in a 14-year-old boy (Wilson et al., 2014). Whilst bioinformatics tools exist for the removal of unwanted reads, the whole genome sequencing of *Leptospira* species

was not possible in this particular case. Beyond species level identification, greater sequencing depth is required to obtain enough target reads to assess important characteristics of an organism such as antimicrobial resistance, virulence and strain identity. In the case of *Ap*, whole genome metagenomic sequencing can be achieved through the bolstering of pathogen abundance through tick cell culture (see Chapter 3) or the collection of infected material at relative bacteraemia peaks in an infected individual (Crosby et al., 2022). The very existence of bacteraemic peaks in non-diseased reservoirs is however debateable, as pathogen loads may not substantially change across the course of infection potentially limiting the number of infected samples susceptible to this approach. As mentioned in chapter 3, establishing tick cell cultures of pertinent strains of *Ap* is challenging, time consuming and crucially, low sensitivity significantly impacting the scope of whole genome sequencing projects. The seeding of strains into cell lines is challenging and often unsuccessful which again limits the number and potentially types of strains that can be studied. Similarly, access to infected animals and identification of bacteraemia peaks pose comparable challenges requiring specific expertise, time and funding, impacting the potential outputs of a project. Microbial DNA enrichment methods offer the potential to bypass these limitations by improving pathogen-host DNA ratios after the collection of infected material. Thus, a larger number of samples can be interrogated with the potential to collect infected tissues from wild animals that may not necessarily be at maximum bacteraemia or accessible through culture techniques. There are examples in other species and genera that illustrate the utility of these methodologies when sequencing difficult to culture organisms. For example, *Candidatus Liveribacter asiaticus* was successfully sequenced from previously inaccessible low titre samples of the bacterium (Cai et al., 2019). Other examples include species such as *Neisseria meningitidis* (Clark et al., 2018) and *Chlamydia trachomatis* (Christiansen et al., 2014) which could be directly sequenced from clinical samples. These case studies not only demonstrate the clinical utility of capture technology but also reveal the intricate biological mechanisms that drive disease and the success of these bacterial lineages.

There are a number of commercialised strategies employed for the enrichment of microbial DNA at every stage of a sequencing project (DNA extraction, post DNA extraction, during library preparation, during sequencing). What is particularly



exciting is the potential to combine enrichment methodologies at all stages to achieve optimal pathogen-host DNA ratios and generate high coverage genomes of difficult to culture bacteria, namely *Ap*. DNA extractions can be used to remove host DNA from the sample whilst retaining all of the microbial DNA (Oney et al., 2021). This typically involves a process known as differential lysis where mammalian cells are lysed initially with a chaotropic buffer, leaving the microbial cells unaffected (McCann & Jordan, 2014). Host DNA is then depleted with chaotrope-resistant DNase molecules prior to the lysis of microbial cells with a suitable lysis reagent. Microbial lysis reagents are typically capable of degrading gram-negative, gram-positive and fungal cell walls (Oney et al., 2021; McCann & Jordan, 2014). This effectively isolates microbial DNA eliminating the host. Post-DNA extraction enrichment kits often exploit the disparity in CpG methylation rates observed between prokaryotic and eukaryotic organisms (Feehery et al., 2013). The NEBNext Microbiome enrichment kit takes advantage of the high CpG methylation rates in eukaryotic host DNA, using a CpG-specific binding protein MBD2 to bind the methylated DNA. Magnetic beads bound to the MBD2 proteins are utilised to separate bound (host) and unbound (microbial) DNA thus enriching for microbial DNA in the sample.

Another promising approach for enriching target DNA sequences is the use of the SureSelect target enrichment system from Agilent. This methodology uses custom-designed biotinylated RNA probes (baits) that are complementary to the DNA sequences of interest (Zeineldin et al., 2023). A DNA hybridisation process is employed to bind target DNA in a sample during library preparation. The biotinylated RNA probes bound to target DNA can be separated from unbound DNA using streptavidin-labelled beads, eliminating unbound host DNA. This methodology has been used to sequence the genomes of a number of microorganisms including *M. tuberculosis*, *Mycobacterium bovis*, *Candidatus Liberibacter asiaticus* and more (Zeineldin et al., 2023; Brown et al., 2015; Doyle et al., 2018; Cai et al., 2019). The final stage of generating genomic data involves the sequencing of DNA molecules in a chosen sequencing system. ONT have developed an in-silico enrichment procedure called Adaptive Sampling (AS) that can be used to enrich or deplete specified DNA sequences. Adaptive sampling requires the user to input a DNA file (BED or fasta format) which will be used to identify reads in real time. In the case of

target enrichment, ~500bp of a strand may be sequenced and basecalled in real time, after which the reference file will be interrogated. If the first 500bp of the read matches the target sequence, the strand will be sequenced entirely; if the strand does not match the reference file, the charge in the pore will be reversed, ejecting the strand, freeing up the pore for a new strand to enter. The fundamental principle of this process is to free up sequencing capacity for target DNA subsequently increasing the share of target DNA in the data. In practise, adaptive sampling has been reported to facilitate a 13.87-fold increases in the least abundant microbial species in a given metagenomic sample (Martin et al., 2022). There are however some caveats, for example, prolonged use of AS has been shown to reduce pore life expectancy and the constant rejection of strands significantly reduces the overall yield of a sequencing experiment (Martin et al., 2022). When these factors are taken into consideration, a 4.93-fold increase was achieved.

The development and optimisation of a financially viable enrichment methodology to sequence *Ap* directly from blood or tissue samples will enable the interrogation of rodent, bird and roe deer associated variants that remain to be sequenced at the time of writing. In addition to this, high-resolution sequencing data from wildlife and livestock that occupy the same geographical areas can better inform farmers on risk factors and determine the reservoir species responsible for spreading veterinary variants of *Ap*. Beyond this, a better understanding of *Ap* diversity across diverse vertebrate hosts may also hold answers to highly relevant questions such as the factors that drive host specificity, and the risk of zoonoses. The effectiveness of each enrichment strategy and all possible combinations of enrichments were explored to determine the limits and best practises when attempting to enrich for *Ap* in sequencing datasets. A variety of bioinformatical techniques will be employed to clean-up, classify, and map reads to determine pathogen host ratios and elucidate how these methodologies translate when constructing draft whole genomes of *Ap* from infected spleen tissues.

## 4.2: Methods

### Sample Collection

23 red deer spleens and 50 roe deer spleens were collected by Forestry England rangers in Grizedale Forest, Cumbria between the 1<sup>st</sup> of November 2022 and 24<sup>th</sup> of March 2023 as part of their nationwide deer management strategy. In addition to this, 80 sheep spleens were collected from a Cumbrian abattoir on the 8<sup>th</sup> of November 2022 to investigate infection rates across farms in Cumbria (Figure 34). Additional details on the number of samples taken from each farm can be found in the appendix (1.0.0). Spleens collected from sheep were dissected by meat process workers and immediately bagged up and placed on ice to limit environmental exposure and degradation. Deer spleens were dissected by Forestry England rangers on site in Grizedale Forest, immediately bagged up and transported to freezers for long term storage. Details of deer age, gender and location was noted.

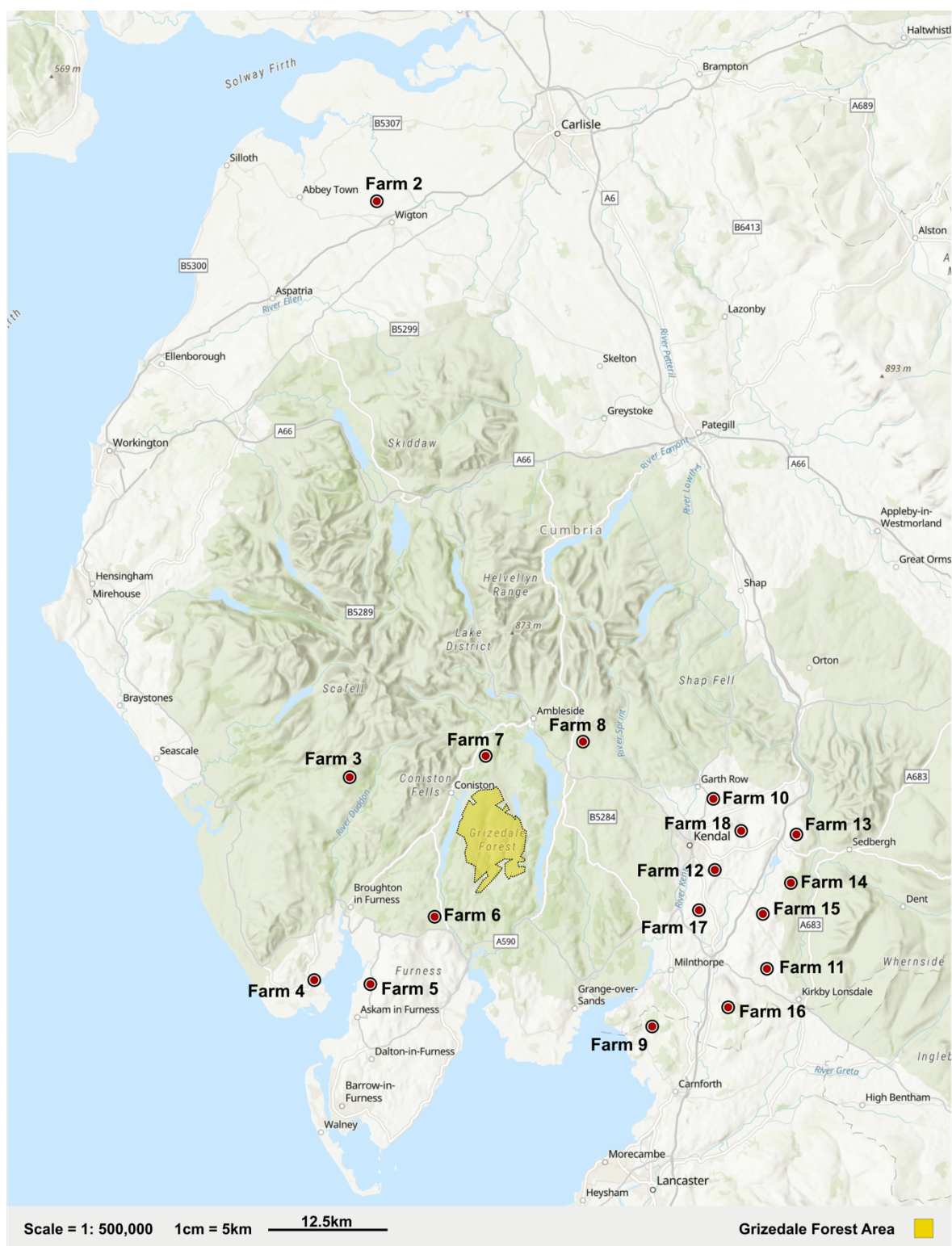
### DNA Extractions

Prior to DNA extractions, all sheep and deer spleens were dissected in order to collect ~5g portions of spleen from the interior of the organ, ensuring aseptic technique and minimal exposure to the environment. Extracted spleen tissues were frozen in 2ml Eppendorf tubes until required for DNA extraction.

### Molzym Complete 5 Differential Lysis Kit (D-321-050)

A total of 1g of spleen tissue was used as starting material for the Molzym Molysis Complete5 protocol for microbial DNA purification. Under aseptic conditions, 1g of tissue was pureed until very smooth using sterile scissors. ~100mg of the pureed spleen tissues was then transferred to a sterile 2.0ml tube and 250µL of chaotropic buffer was added. All subsequent steps were performed according to manufacturers' instructions. Upon completion of the protocol. DNA quality, quantity and molecular weight were assessed with three platforms, Thermo Fisher Scientific Nanodrop One

(ND-ONE-W), Thermo Fisher Scientific Qubit 3.0 fluorometer (Q33216) and Agilent Tapestation 2200 respectively.



**Figure 34: Map of Cumbria, UK with the locations of farms which provided sheep for slaughter at the abattoir on the 8<sup>th</sup> of November 2022. The location of Grizedale Forest is also indicated where deer spleens were collected from the 1<sup>st</sup> of November 2022 to 24<sup>th</sup> March 2023.**

### NEB Monarch HMW DNA Extraction Kit (T3060S/L)

Monarch HMW DNA extractions were performed on all samples following the manufacturers guidelines for the T3060L Monarch HMW DNA Extraction from Tissue kit. DNA quality, quantity and molecular weight were assessed with three platforms, Nanodrop, Qubit and Tapestation respectively.

### NEB Microbiome Enrichment Kit (E2612S/L)

An input of 1ug DNA in 20µL volumes was used for each NEB microbiome enrichment. MBD2-Fc-bound magnetic beads were prepared according to manufacturer's instructions. Enriched microbial DNA was captured and purified using Agencourt AMPure XP beads following option A guidelines in the NEB microbiome enrichment kit manual. Purified enrichments were verified using the Thermo Fisher Scientific Qubit 3.0 fluorometer (Q33216) to assess quantity and the Agilent TapeStation 2200 to assess molecular weight using either D1000 (5067-5582) or genomic (5067-5365) screentape and their respective reagents (5057-5586, 5067-5583; 5067-5365).

### Real-time PCR

A previously described real-time PCR assay (Courtney et al., 2004) was utilised for the high sensitivity detection of *Ap* using primers: *ApMSP2f* (5'-ATGGAAGGTAGTGTGGTTATGGTATT), *ApMSP2r* (5'-TTGGTCTTGAAGCGCTCGTA). This generated a 77bp fragment using TaqMan probe *ApMSP2p-FAM/BHQ*: (5'-TGGTGCCAGGGTTGAGCTTGAGATTG) Which is dual labelled with FAM/BHQ. The PCR was performed with a reaction volume of 25µL using the Brilliant Quantitative PCR core reagent kit with SureStart Taq DNA polymerase in a Bio-Rad Opticon thermal cycler. Reaction conditions were as follows: primers were concentrated at 900nM each, the probe, *ApMSP2p-FAM/BHQ* at 125nM, with 2µL of template DNA. Cycling conditions included an initial activation of the Taq DNA polymerase for 10 minutes at 95°C, followed by 40 cycles of a 15 second denaturation at 95°C, followed by a 1-minute annealing extension step at 60°C.

## Statistics

A one-factor analysis of variance (ANOVA) was performed in IBM SPSS Statistics (v29) to assess differences in infection intensity between species, Ct values from qPCR as the dependent variable. Species was treated as the independent factor. The data were tested for normality using the Shapiro–Wilk test and for homogeneity of variance using Levene’s test. A significance threshold of  $P < 0.05$  was applied.

## Agilent SureSelect XT HS2 Platform

Probes were designed by Agilent Technologies staff on proprietary software to be complementary to >10Mbps of *Ap* genomic data derived from the seven UK strains and German strain RD1. Probes ranged in size from 120bp to 150bp in length. A tier 2 package was purchased meaning ~140k different probes were designed, printed and pooled.

Prior to Library preparation, all samples were fragmented using the Diagenode Bioruptor Pico (Cat# B01080010). Enriched DNA was diluted into 50µL containing ~400ng DNA in 1.5ml Eppendorf tubes. The Bioruptor Pico was prepared according to manufacturer’s instructions, cooling fresh purified water to 4°C. DNA fragmentation for fragments ranging in size from 500bp - 700bp was achieved using 2 cycles of 25 seconds ON, 30 seconds OFF ensuring samples were placed on ice prior to fragmentation for 15 minutes. Once fragmentation was complete samples were spun at 300rpm for 1 minute. Fragmentation was confirmed on the Agilent Tapestation 4150 (G2992AA) using D1000 high sensitivity screen tape (5067-5582) and reagents (5057-5586, 5067-5583) according to the manufacturer’s guidelines.

The Agilent SureSelect XT HS2 DNA System protocol (Version E0, July 2022) was followed to prepare eight libraries for the eight pilot samples (Sheep: WR6, AEH3; Red Deer: GR01, GR03; Roe Deer: GRD08, GRD18, GRD09, GRD17). Notably, DNA quantity was doubled to 400ng per library and hybridisation was performed overnight to improve capture efficiency. A total of 20 cycles was required to amplify post-capture libraries back up to suitable concentration for sequencing whilst minimising duplication.

The MiSeq V3 600-cycle kit (MS-102-3003) was utilised to sequence all eight libraries. The reagent kit was thawed and prepared according to manufacturer's instructions. Library concentrations were calculated with the Thermo Fisher Scientific Qubit 3.0 fluorometer (Q33216) with the high sensitivity assay kit (Q32851) and the Agilent Tapestation 4150 (G2992AA) with D1000 screentape (5067-5582) and reagents (5057-5586, 5067-5583). Libraries were then pooled into a single equimolar library, denatured and diluted to 12.5pM prior to loading into the V3 reagent cartridge.

### Adaptive Sampling

Long read sequencing was performed with the PromethION P2 solo using ONT's V14 chemistry and a PromethION R10.4.1 flow cell. Enriched DNA libraries were sequenced using the native barcoding kit 24 V14 (SQK-NBD114.24) according to the 'ligation sequencing gDNA – native barcoding kit 24 V14' protocol available on ONT's website. Libraries were quantified using the Thermo Fisher Scientific Qubit 3.0 fluorometer (Q33216) and the Agilent tapestation 4150 (G2992AA) using genomic screentape (5067-5365) and reagents (5067-5365). AS was activated in minKNOW in the "start experiment" options menu specifying the enrichment of DNA strands that match a provided fasta file (Harris Genome). Basecalling was set to "high accuracy" using either the Guppy or Dorado integrated basecaller which was dependent of the version of minKNOW operating the P2 solo. The system was then run for 24 hours before the flow cells were washed and reloaded with excess library using the wash kit (EXP-WSH004-XL) following the manufacturer's guidelines. The libraries were then run for a further 48 hours. Reads were deposited in fastq format and concatenated into single files for each barcode in preparation for processing and analysis.

---

## **4.3: Results**

### Infection Rates & Intensity

A total of 23 red deer spleens, 50 roe deer spleens and 80 sheep spleens were tested for *Ap* infections. The infection rates in red deer were 91.3% (21/23), in roe deer 88% (45/50), and in sheep 10% (8/80). The qPCR cycle threshold (Ct) values obtained from the positive samples varied between 30 and 18.8, with average Ct



values of 27.15 for sheep, 24.68 for red deer, and 25.67 for roe deer (Figure 35). Although red deer on average had lower Ct values, only roe deer were found to have Ct values lower than 20. A one factor Anova comparing the mean Ct of each species generated a P-value of 0.363, indicating there is no significant difference in infection intensity between species. Infection intensity remained relatively stable throughout the collection of the spleens but an overall upwards trend in Ct value from November 2022 to March 2023 was observed. A total of 60 female and 12 male deer were

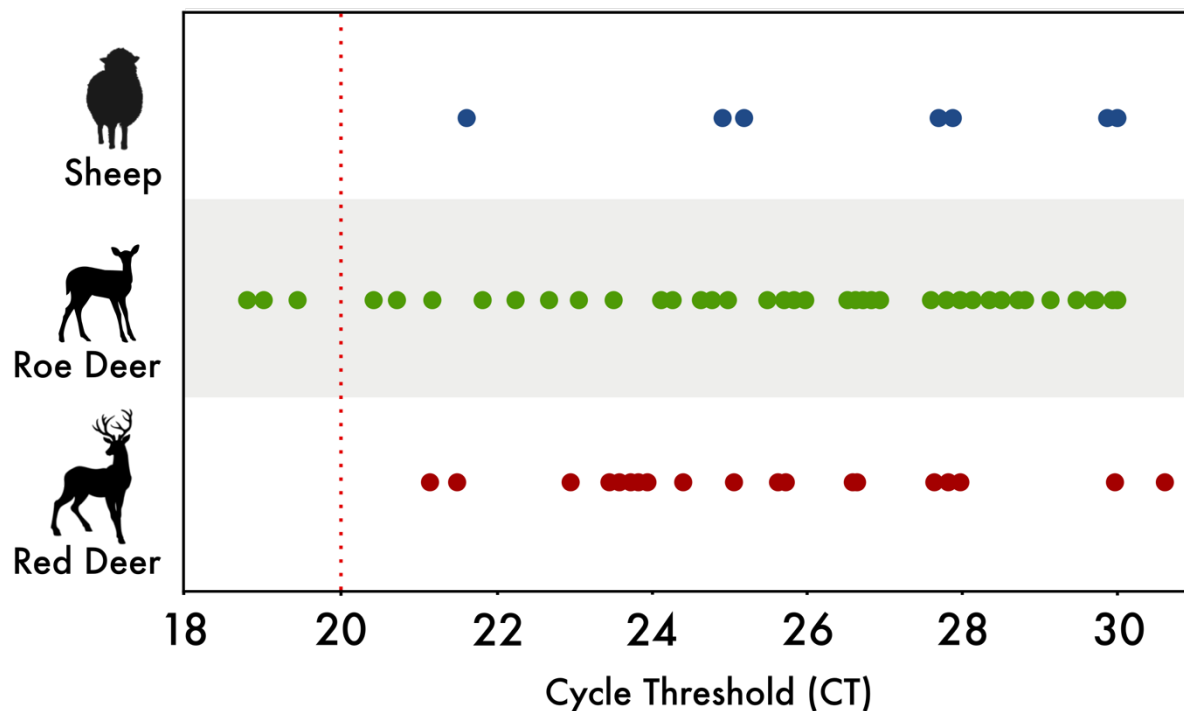


Figure 35: Relative infection intensity of *Anaplasma phagocytophilum* in red deer, roe deer and sheep collected in Cumbria, UK between November 2022 and March 2023. Measured with a Bio-Rad Opticon using a real-time PCR assay targeting a 77bp fragment of msp2

investigated, 42 of which were adults, 21 juveniles and 9 yearlings. Gender and age of the wild deer had no impact on infection intensity. Positive sheep spleens originated from five farms (3, 5, 6, 7 & 17). Infection rates were as follows: 25% (2/8), 12.5% (1/8), 100% (3/3), 25% (1/4), and 16.7% (1/6) respectively. Sampling sizes for each farm was limited to the number of sheep sent to slaughter (between 1 and 8 per farm). Raw data is available in the appendix (1.0.0) with information relating to date of collection, PCR results, age and farm of origin. Additionally, details of deer collected can be found in the appendix (1.1.0) with date of collection, PCR result, age and gender details.



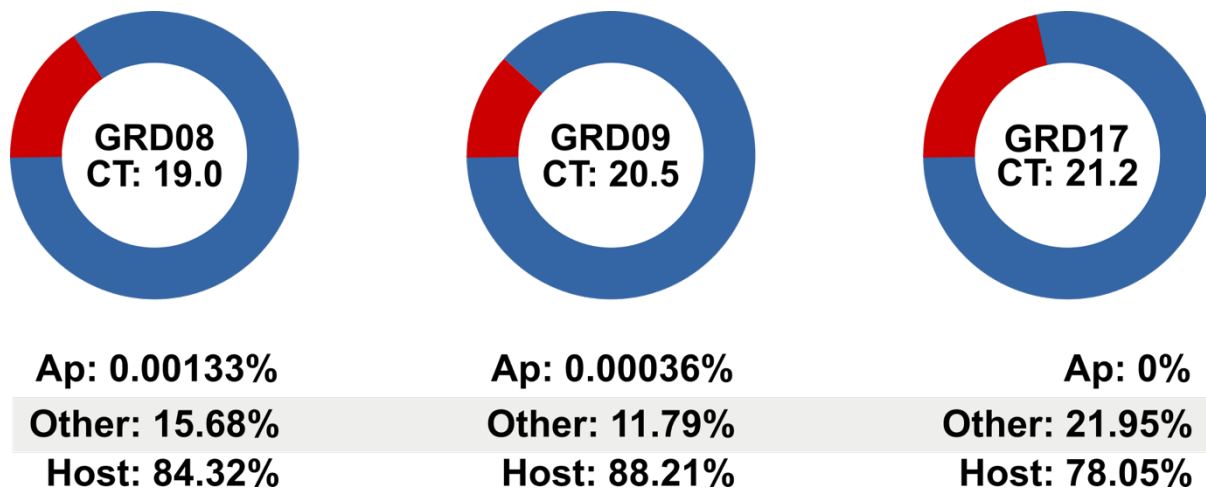


Figure 36: Benchmark sequencing efficiency achieved on the PromethION P2 solo utilising the NEB Monarch HMW DNA extraction kit for tissue on infected roe deer spleens (GRD08, GRD09, GRD17). Cycle threshold (Ct) is indicated for all three samples ranging from 21.2 - 19.0. Relative population sizes of host, other and *Anaplasma phagocytophilum* (Ap) are estimated using Kraken 2 read classification data generated against a custom database

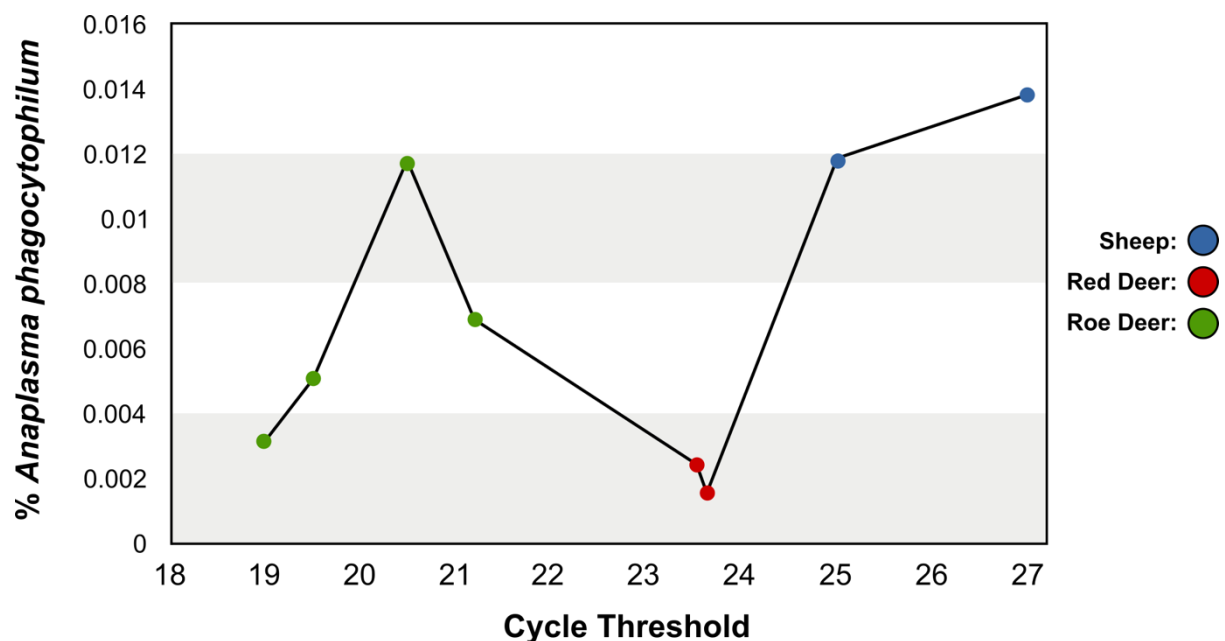
### Benchmarking

Three roe deer samples were taken forward for evaluation after considering Ct values and DNA concentrations. These samples were GRD08, GRD09 and GRD17, which yielded Ct values of 19, 20.5 and 21.2 respectively. The samples were sequenced with the PromethION P2 solo and the generated reads were classified using a custom Kraken 2 database. The analysis revealed that both the (Ap: host) and (Ap: non-target pathogen) ratios were unfavourable for generating whole genome assemblies of Ap, with just 0.00133%, 0.00036% and 0% of reads mapping respectively (Figure 36). Host reads made up the majority of the data generated ranging from 78.05% to 84.32%. Interestingly, the microbial population in GRD17 represented a significant proportion of reads with 21.95% identifying with other microbial organisms present in the deer's spleen or introduced shortly after collection from the environment (Figure 36). The primary microbial populations detected belonged to *Clostridium* sp. and *Babesia* sp.

### Adaptive Sampling

After the establishment of sequencing benchmarks, the impact of AS, Ct, and host identity were investigated with a separate run on the PromethION P2 solo. Eight samples were chosen for this run due to their spectrum of Ct values and origins in different mammalian hosts (4x roe deer, 2x red deer and 2x sheep). In ascending order of Ct, the samples chosen were: GRD08, GRD18, GRD09, GRD17, GR01,

GR03, WR6 and AEH3. The Harris genome evaluated in chapter 3 was used as an enrichment target for minKNOW. Contrary to expectations, the enrichment appeared to be most successful in the sheep derived DNA samples with the proportion of *Ap* reads reaching 0.014% despite lower infection intensities (Figure 37). Of the previously benchmarked samples GRD08 and GRD09, AS improved the proportion of *Ap* reads by 2.3x, 32.8x respectively. In the case of GRD17, 0.00690% of reads were classified as *Ap* representing the first data generated for this sample as the benchmarking run failed to detect *Ap*. The majority of reads generated were short (<1kb) likely impacting the final yield of the experiment. Upon inspection of the “read-end” reason for each read, it is clear that adaptive sampling was in action as the majority of these short reads had been terminated with a voltage reversal.



**Figure 37:** The relationship between cycle threshold (infection intensity) as measured by a real-time PCR assay targeting a 77bp fragment of the *MSP2* gene and the percentage of reads identified as *Anaplasma phagocytophilum* when sequenced with the PromethION P2 solo with adaptive sampling active and enriching for the Harris genome.

## Bait Capture

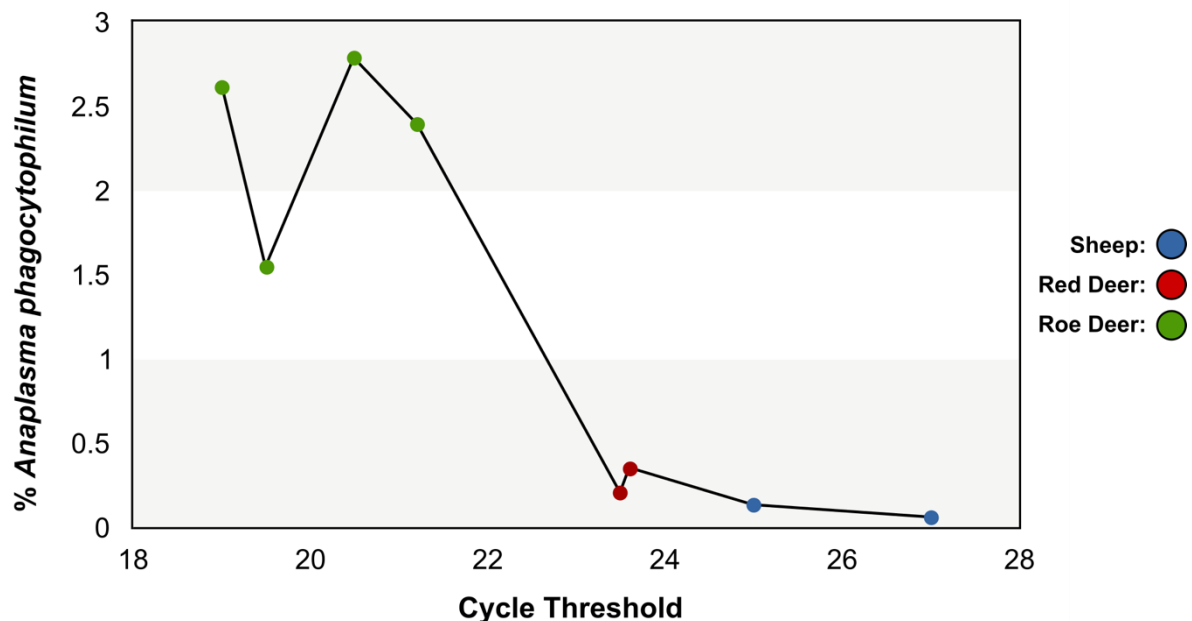


Figure 38: The relationship between cycle threshold (infection intensity) as measured by a real-time PCR assay targeting a 77bp fragment of the MSP2 gene and the percentage of reads identified as *Anaplasma phagocytophilum* after bait capture and sequencing with the Agilent SureSelect platform and Illumina MiSeq V3 respectively.

The same set of samples were subjected to sequencing on the Illumina MiSeq using the SureSelect workflow designed by Agilent which uses custom biotinylated baits. The baits were designed to be complementary to >10MB of *Ap* genome sequences spanning seven genomes from the UK sequenced in chapter 3 and the RD1 genome derived from a roe deer in Germany. Due to the repetitive nature of the genomes a tier I probe package was sufficient to cover all parts of the genome. The baits performed more predictably than AS, with Ct largely predicting the efficiency of the capture steps (Figure 38). Bait capture outperformed AS for all samples ranging from a 4.3x enrichment in AEH3 to an 837.1x enrichment in GRD08 when directly comparing both methodologies. When considering the benchmark sequencing for GRD08 and GRD09, enrichments of 1963.3x and 7685.6x were achieved respectively.

## Combined Enrichments

In my efforts to enhance the representation of *Ap* in sequencing data, I explored the use of AS or bait capture alongside two upstream preparation methods: the Molzym Complete 5 differential lysis kit and the NEB microbiome enrichment kit. We replaced the Monarch HMW DNA extraction kit with the Molzym Complete 5 kit, while the NEB

microbiome kit was used post DNA extraction and therefore compatible with both methods. I conducted three sequencing experiments (2x ONT, 1x Illumina), covering all possible enrichment combinations, including bait capture or AS (Figures 39 and 40)

Across all enrichment conditions, bait capture on the Illumina MiSeq consistently outperformed AS on the PromethION P2 solo. Notably, the combination of the Monarch HMW DNA extraction and NEB microbiome enrichment provided the most favourable pathogen:host ratios for target enrichment with SureSelect baits, achieving a maximum share of 65.74% reads assigned to *Ap* in GRD08, representing an enrichment of 49454x enrichment on the benchmark for this sample. The Ct values proved to be a significant predictor of sequencing success, with an increase of 2.1 cycles reducing the share of *Ap* reads to 15.39% under optimal conditions.

However, the Molzym DNA extraction kit performed poorly when combined with the NEB microbiome enrichment kit, resulting in shares of 6.35% (GRD08), 1.30% (GRD09), and 1.03% (GRD17). When used alone, the Molzym DNA extraction kit produced read shares of 10.69%, 0.48%, and 0.44%, respectively.

The proportion of *Ap* in the sequencing data remained consistent with the Illumina data when using the P2 solo, regardless of whether AS was active or not. The combination of the NEB microbiome enrichment kit with the Monarch HMW DNA extraction kit yielded the highest proportion of desired reads. However, despite notable enhancements observed with active adaptive sampling in prior experiments, this technology failed to deliver definitive improvements in this subsequent experiment. In fact, it resulted in lower yields across four samples compared to sequencing without AS activation.

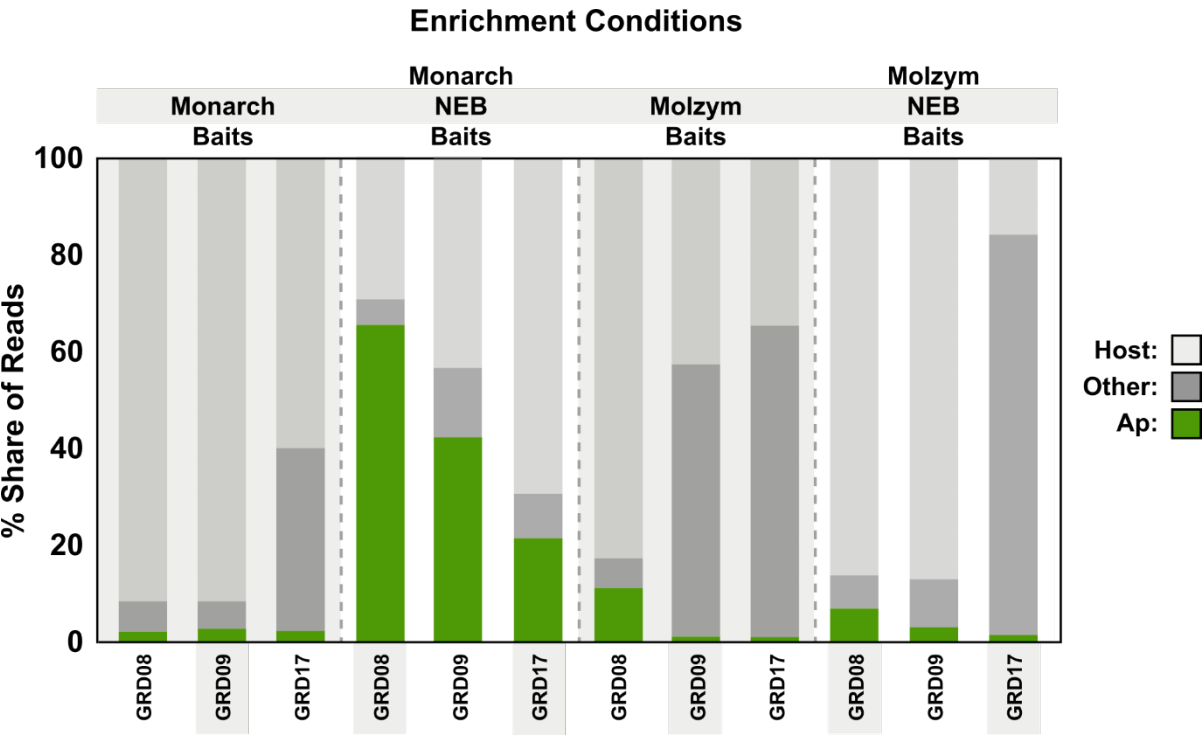


Figure 39: The percentage share of *Anaplasma phagocytophilum* (Ap) reads in a sequencing dataset of 3 samples (GRD08, GRD09, GRD17) subjected to four different enrichment techniques: (1) Monarch HMW DNA extraction and SureSelect bait capture; (2) Monarch HMW DNA extraction, NEB microbiome enrichment, SureSelect bait capture; (3) Molzym Complete5 extraction and SureSelect bait capture; (4) Molzym Complete5 extraction, NEB microbiome enrichment and SureSelect bait capture. Green bars (Ap) represent reads classified as *Anaplasma phagocytophilum* (Ap), dark grey bars (Other) represent reads classified as non-target microbes, light grey bars (Host) represent reads classified as roe deer. Classification of reads was performed with Kraken 2 using a custom database.

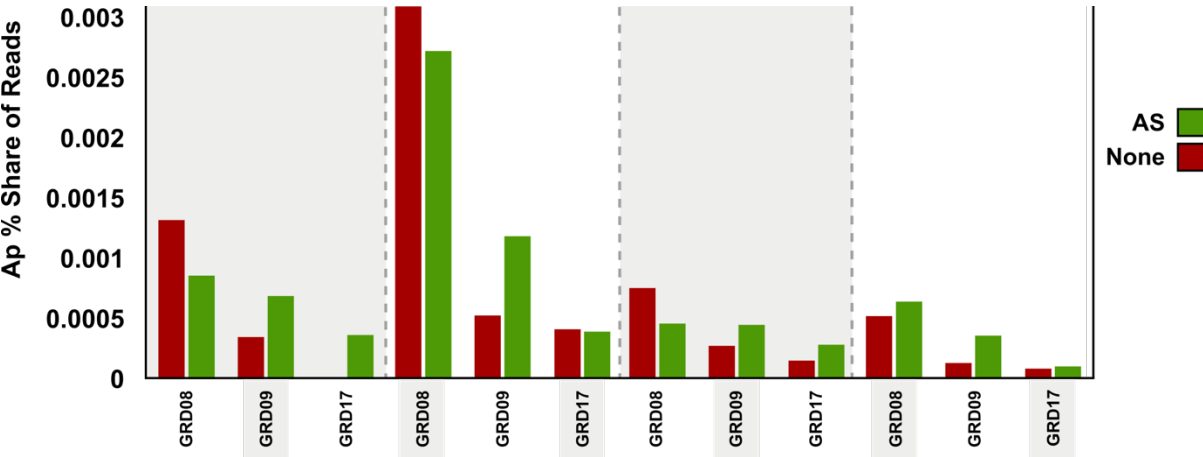


Figure 40: The percentage share of *Anaplasma phagocytophilum* (Ap) reads in a sequencing dataset of 3 samples (GRD08, GRD09, GRD17) subjected to four different enrichment techniques with (AS, green) or without (None, red) adaptive sampling: (1) Monarch HMW DNA extraction (2) Monarch HMW DNA extraction and NEB microbiome enrichment; (3) Molzym Complete5 extraction; (4) Molzym Complete5 extraction and NEB microbiome enrichment. Classification of reads was performed with Kraken 2 using a

In summary, our investigation into optimising Ap representation in sequencing data elucidated several key findings. Bait capture on the Illumina MiSeq platform consistently outperformed AS on the PromethION P2 solo, with the combination of

the Monarch HMW DNA extraction and NEB microbiome enrichment yielding the highest proportion of desired reads. Despite the promising potential of active adaptive sampling observed in prior experiments (Figure 37), its implementation in the follow up experiments (Figure 40) failed to demonstrate definitive improvements. These results underscore the intricate interplay between upstream preparation methods, enrichment techniques, and sequencing platforms in achieving optimal pathogen detection.

### Quantifying Genome Coverage

In previous experiments (Figures 41, 42, 43), the focus was on improving enrichment strategies to quantify the presence of *Ap* reads in sequencing data. However, it was equally important to examine how these reads aligned to a reference genome and whether this method could be extended to assemble whole genomes from metagenomic data. To address this, the next step involved aligning the reads to a UK-derived reference genome, (Harris) sequenced in Chapter 3. After alignment, I used a custom Python script to investigate gaps in coverage and analysed their relationship with GC content using BAM2plot. These analyses provided a more detailed understanding of *Ap* read alignment patterns and the factors that could influence genome assembly. Our findings revealed an interesting trend in the relationship between the number of reads and the proportion of the reference genome they covered. Initially, as the number of reads increased, so did genome coverage. However, after a certain point, additional reads yielded diminishing returns in overall genome coverage. Despite this plateau, the proportion of the genome with at least 30x coverage continued to increase with more reads. Peaks in coverage were especially prominent in repetitive portions of the *msh2* gene, indicating a focal point for read enrichment and alignment. Although gaps in coverage existed, they were generally small, suggesting that the majority of the reference genome was captured effectively by the sequencing reads. This pattern highlights the success of our enrichment methodologies in covering the whole genome, even with occasional gaps in coverage.

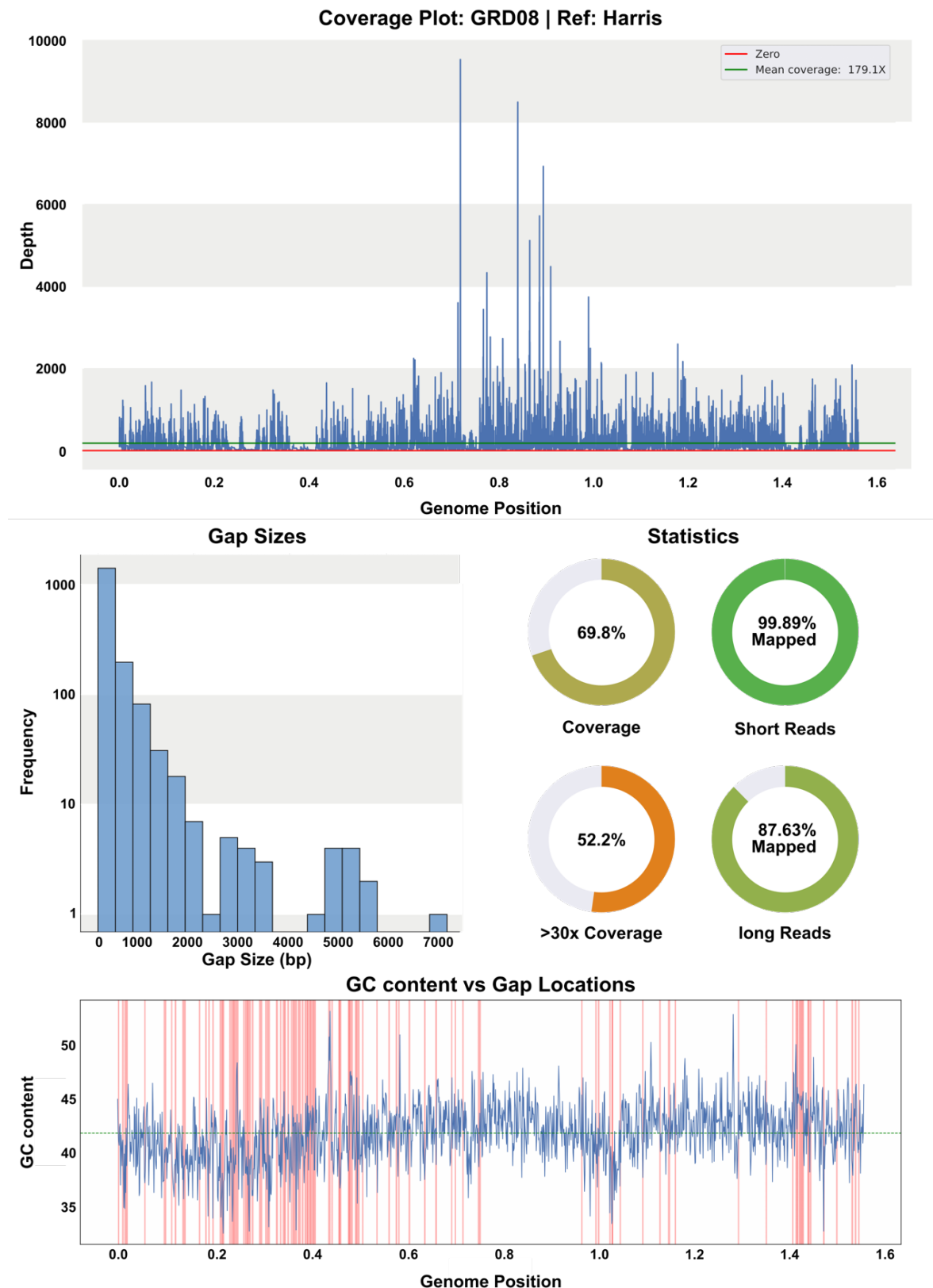


Figure 41: Coverage statistics for reads generated for a roe deer derived *Anaplasma phagocytophilum* (Ap) strain GRD08 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository.

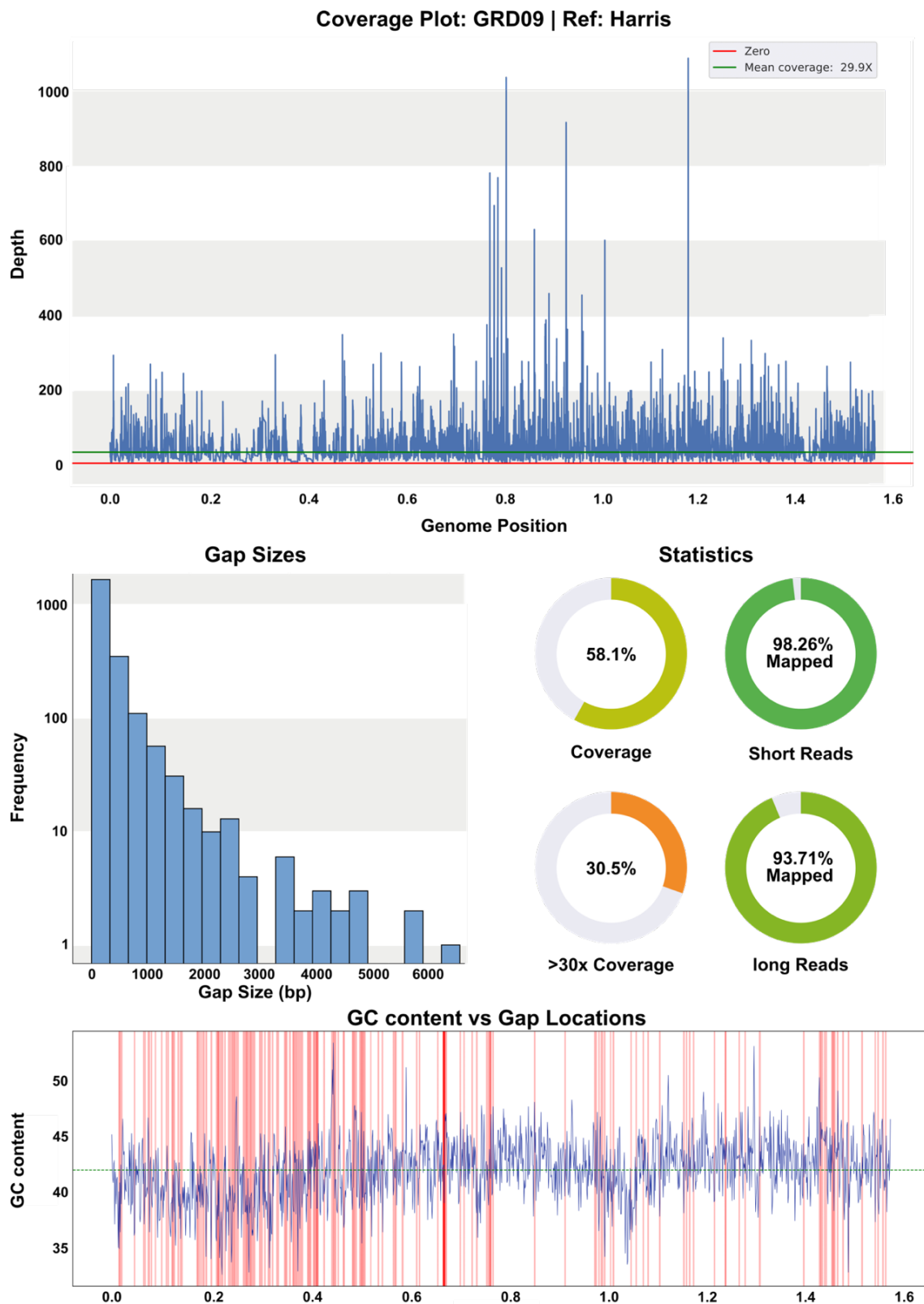
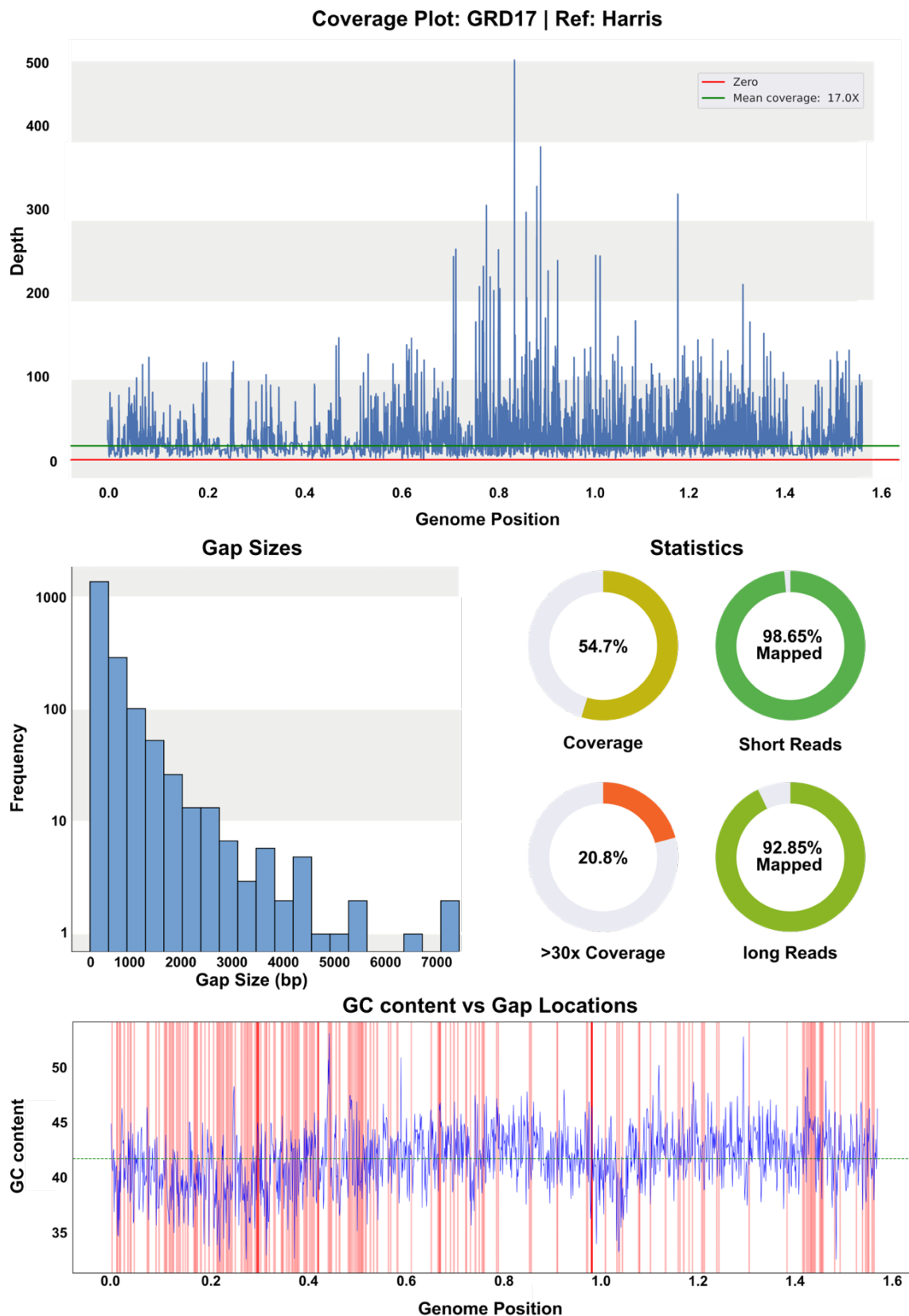


Figure 42: Coverage statistics for reads generated for a roe deer derived *Anaplasma phagocytophilum* (Ap) strain GRD09 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository.





**Figure 43: Coverage statistics for reads generated for a roe deer derived *Anaplasma phagocytophilum* (Ap) strain GRD17 against a reference UK derived Ap strain, Harris. Top plot was generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be Ap group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository.**

## 4.4: Discussion

*Ap* is a highly diverse obligate intracellular bacterium that comprises various epidemiologically distinct ecotypes. Our current understanding of its ecology and epidemiology stems primarily from spatial and temporal studies that rely on a single genetic locus to detect and characterise *Ap* variants across a range of host and vector species (Malmsten et al., 2014; Tolnai et al., 2015; Jahfari et al., 2014; Gandy et al., 2022). Whole genome data on the other hand is limited to a handful of genomes of varying quality and completeness (Table 2). Analysis of these whole genomes shows that the full extent of *Ap*'s genetic variation is far from captured, with most data coming from the clonal USA *Ap*-HA group or European ecotype I. This limited dataset restricts the conclusions that can be drawn, particularly in regions like Europe and Asia, where higher genetic diversity is expected (Jahfari et al., 2014; Gandy et al., 2022; Mukhacheva et al., 2019). The scarcity of comprehensive genome data is largely due to the difficulty in culturing *Ap* and the low sensitivity of traditional culturing methods (Dugat et al., 2015). Current sequencing methods often require a culture step to achieve a high pathogen-to-host ratio, which allows for better representation of *Ap* in sequencing datasets. However, enrichment techniques used for closely related species, like *Anaplasma platys*, have successfully bypassed this culture requirement, enabling direct sequencing from infected dog blood (Llanes et al., 2020). Despite these advancements, no studies have yet applied this technique to sequence *Ap* directly from wild reservoir hosts using spleen tissue. This chapter investigates the challenges of assembling *Ap* MAGs from wild animal tissues and evaluates optimal methods for overcoming these obstacles.

Initially various reservoir hosts including sheep, red deer and roe deer were investigated for the presence of *Ap* infections using a highly sensitive real-time assay targeting the numerous *msp2* pseudogenes present in all strains of *Ap* (Courtney et al., 2004). The infection rates amongst these species were notably varied, with red deer exhibiting a rate of 91.3% (21/23), roe deer 90% (45/50), and sheep 10%. The high infection rates amongst deer in Grizedale Forest solidifies their place as important reservoirs hosts of *Ap* in the UK and beyond and corroborates previous studies conducted in Europe that have found infection rates between 10-88% in red deer and 10-99% in roe deer (Stuen et al., 2013). As for previous UK based studies,

10-20% of roe deer and 80% of red deer were found to be infected (Robinson et al., 2009). Despite red deer displaying lower average Ct values, only roe deer had Ct scores lower than 20. Statistical analysis through one-factor ANOVA indicated no significant difference in infection intensity between species implying each host species was not significantly more susceptible to high intensity infections. There are a number of possible explanations for these results beyond the limited sample sizes investigated. Bacteria showing no variation in infection intensity across multiple host species can be the result of factors such as broad host adaptation, conserved virulence mechanisms, similar host susceptibility, or generalist infection strategies (Keebaugh et al., 2014). European ecotype I is considered a host generalist variant of *Ap* (Jahfari et al., 2014), and the most commonly encountered variant in the UK (Gandy et al., 2022). Therefore, the strains present in Grizedale Forest deer and Cumbrian sheep may fall into this host generalist category, a hypothesis that could be confirmed in future investigations through the amplification and sequencing of *groEL* in the *Ap* positive DNA samples.

Interestingly, infection intensity remained relatively stable across the spleen collection period, although an overall upward trend in Ct values was observed from November 2022 to March 2023, potentially the result of increased tick activity as weather conditions improved. Gender and age of the wild deer did not appear to impact infection intensity. An infection rate of 10% in sheep may be cause for concern for farmers, however, there is no previous surveys identifying the potential impacts of *Ap* on farms in this location and how these infections in practise can generate rising costs. Positive sheep spleens originated from five distinct farms, primarily those situated in fells near Grizedale Forest suggesting upland areas in close proximity to deer populations encompasses a major risk factor when controlling for the disease. Previous studies have explored the role of deer in maintaining and spreading *Ap*. Variations in the density of questing *I. ricinus* ticks can be directly correlated with deer abundance due to the high tick burden species of deer experience (Ruiz-Fons & Gilbert, 2010). It is therefore feasible that farms in close proximity to Grizedale Forest are experiencing not only higher tick densities but higher rates of infection amongst questing ticks due to feeding on infected deer. The North of England in particular appears to be a hotspot in the UK for *Ap* infection in tick populations with ecotype I accounting for the majority of infections (Gandy et al.,

2022). Some studies have however shown that *Ap* can be maintained in populations of sheep in the absence of deer, implying proximity to deer and the potential knock-on effects are not the whole explanation (Stuen et al., 2013). By developing whole genome sequencing techniques to analyse *Ap* directly from spleen tissue, researchers could accurately fingerprint strains. This advancement would significantly improve our understanding of strain distribution within deer and sheep populations, helping to better assess the potential risks posed by the interaction of wild deer and livestock in shared habitats.

Three roe deer samples, GRD08, GRD09, and GRD17, with Ct values ranging from 19 to 21.2, were selected for evaluation based on Ct and DNA availability. Utilising the PromethION P2 solo platform, the samples underwent sequencing, and subsequent classification using a custom Kraken 2 database. Unsurprisingly, the analysis revealed unfavourable ratios for generating whole genome assemblies of *Ap*, with only 0.00133%, 0.00036%, and 0% of reads mapping to *Ap* in GRD08, GRD09, and GRD17, respectively. Host reads predominated, comprising 78.05% to 84.32% of the generated data. Notably, GRD17 exhibited a significant microbial population, with 21.95% of reads identifying with other microbial organisms found in the deer's cardiovascular system or introduced post-collection. The primary microbial populations detected included *Clostridium sp.* and *Babesia sp.* The presence of *Clostridium sp.* was likely down to environmental contamination during the collection by forestry England rangers. The large population of *Babesia sp.* in roe deer implies a reservoir role, consistent with surveys which find roe deer to be the main vertebrate host for species such as *B. venatorum* in Europe (Michel et al., 2014). To increase the presence of *Ap* in our sequencing data, we employed two main strategies: adaptive sampling (AS) and bait capture. These technologies allowed us to selectively enrich for *Ap* DNA while minimising the background noise from host and non-target DNA. Our initial experiments provided positive results with improvements in both raw read counts and percentage share of reads that could be classified as *Ap*. Bait capture proved to be the most effective methodology surpassing AS for all samples tested. AS appeared to perform particularly well for the sheep samples AEH3 and WR6 which despite having the highest Ct scores had the two highest shares of *Ap* reads in their datasets. I can think of two possible explanations for this, AS works particularly well for low abundance samples, or the

strains present in the sheep share higher homology with the reference genome (Harris) used to guide the software. Similar observations have been made by Martin et al., 2022 finding 13.87-fold enrichments of the lowest abundance species. As for target specificity there is limited documentation regarding the impact of sequence homology, therefore future experiments could explore the limits of AS when targeting divergent strains. Therefore, infection intensity may not be as crucial for success when running AS, instead, similarity of the reference file may be the key metric for success. Ct proved to be a more reliable predictor of success when employing bait capture, however, strain identity will undoubtedly play a key role in the effectiveness of the hybridisation of target DNA and removal of non-specific strands. The data from these experiments was combined and aligned to the Harris reference genome. Alignment of reads revealed that over 50% of the GRD08 genome had been captured across both sequencing runs. With the availability of additional library for GRD08, a second run was conducted on the Illumina MiSeq and SureSelect platform with the aim of completing the GRD08 genome. This increased the number of reads by 1000% but when aligned back onto the reference, coverage increased by just 5% for a total of 55%. This result clearly indicates that during the capture step of the Agilent SureSelect protocol, we were unable to capture all portions of the *Ap* genome. This alerted to potential issues with capturing certain portions of the genome, whilst also indicating that multiple library preparations and target capture steps may improve the proportion of the genome captured. GC-rich regions, low complexity regions, and repetitive elements have all been implicated as potential barriers to the complete capture of an *Ap* genome. Evaluations of the Agilent SureSelect platform have already identified low representation of GC-rich segments (Bodi et al., 2013; Adams et al., 2010). When combined with sequencing bias on the Illumina platform, this may impose significant limitations on the successful capture of the *Ap* genome which typically has a GC content of 41% - 43%.

Additionally, we utilised specific kits for sample preparation before sequencing, such as the Molzym complete 5 differential lysis kit and NEB microbiome enrichment kit, to optimise the quality and quantity of *Ap* DNA present in the extractions. These methods were systematically evaluated using both the PromethION P2 solo and Illumina MiSeq platforms to determine their effectiveness in enhancing the representation of *Ap* in our sequencing datasets. We restricted this analysis to three

samples GRD08, GRD09 and GRD17 due to costs and the availability of DNA. The NEB microbiome enrichment kit combined with the Monarch HMW DNA extraction kit generated optimal results on both the SureSelect platform and on the P2 Solo with or without AS. Combining the Molzym Complete 5 DNA extraction kit and the NEB microbiome enrichment kit produced notably worse results, indicating the kits lack compatibility. The primary reason for poor compatibility is the lower strand lengths preserved by the Molzym extraction process. For effective depletion of methylated DNA with the NEB microbiome enrichment kit, HMW DNA is essential (Yigit et al., 2016). Enrichment of HMW DNA extractions with the NEB microbiome enrichment yields a higher fraction of *Ap* reads and improved read quality resulting in a reduced cost of downstream data generation and analysis, making the methodology viable for larger scale sequencing projects and clinical identification of *Ap* strains. Due to the extremely low populations of *Ap* within samples, methylation depletion followed by hybridisation capture with SureSelect baits significantly reduced the concentrations of DNA in library preparations. In some cases, over 20 PCR cycles were required to bring concentrations back to suitable levels for sequencing. This undoubtedly increases the risk of duplication errors (Zeineldin et al., 2023). Unfortunately, AS appeared to provide little to no benefit when employed to target *Ap* during the enrichment evaluation. There were four cases of a reduced proportion of *Ap* reads when compared to sequencing without the *in-silico* enrichment running (Figure 40); including the optimal Monarch DNA extraction and NEB microbiome enrichment combination. This suggests that multiplexing more than eight samples may impact the performance of the basecalling and identification of *Ap* DNA strands. Nevertheless we provide some evidence that AS holds promise to become a cost-effective method of enrichment (Figure 37) as more advanced algorithms are introduced that can streamline the decision process with more accurate strand rejection (Lin et al., 2024).

I developed several custom Python scripts to support the analysis of genome coverage, focusing on identifying potential capture biases and gap sizes in genome assemblies aligned to the Harris reference genome. These tools are available on GitHub under the name **GapPlotter** (<https://github.com/Wizical/GapPlotter>). The scripts provide functionality to analyse genome coverage for each of the three roe deer samples, starting with a BAM file parser that extracts gap information. This

parser outputs a CSV file containing the start and end coordinates of gaps, as well as the gap lengths. The CSV file is then used to create a histogram of gap lengths and a GC content plot, highlighting the positions of gaps. Additionally, I employed BAM2plot to generate coverage plots, helping to identify regions with unusually high or low coverage across the reference genome. Through these analyses, it became evident that the biotinylated baits used in this study showed a bias towards capturing repetitive regions of the genome, such as the *MSP2* gene, with over 9000x coverage in sample GRD08. Capture biases are common when using newly designed baits (Perry et al., 2010), and future experiments could benefit from refining bait design to target underrepresented regions, thereby improving coverage of low or no coverage areas. Ultimately, the combined sequencing data from this study allowed us to capture at least half of the genome for all three samples. Gap sizes ranged from 7 bp to 7,437 bp, with an even distribution of gaps across the genome. Factors such as GC content, low sequence complexity, repetitive elements, and the conditions of the capture technology (Ct values) all contributed to the capture efficiency.

Going forward, the adaptation of the SureSelect platform to ONT systems may drastically increase genome coverage if sufficient read lengths can be achieved. Using the gap data, we can calculate that a maximum read length of 6kb would be sufficient to capture the entirety of the *Ap* genomes present in GRD08, GRD09 and GRD17 given the current anchor points captured by the baits are still viable.

# CHAPTER 5

## 5.0: Ecotype Resolution in *Anaplasma phagocytophilum* Applying Enrichment Strategies to a Global Context

### 5.1: Introduction

*Ap* is sustained through enzootic cycles involving both tick vectors and vertebrate reservoir hosts (Jahfari et al., 2014). Despite being classified as a single species (Chastagner et al., 2017), *Ap* demonstrates a notable ability for host specialization. This is evident as isolates from different host species do not consistently infect other hosts outside their primary ecological niche (Jahfari et al., 2014). Through molecular ecotyping, researchers have categorised *Ap* into several clusters and subclusters (Jahfari et al., 2014; Huhn et al., 2014). Ecotypes refer to populations occupying distinct ecological niches, and the delineation of these strains relies on molecular data and observed host preferences. The *groEL* gene has proven particularly useful in defining these ecotypes, offering better resolution than more conserved markers like 16S rRNA (Jahfari et al., 2014). A pivotal study by Jahfari et al. 2014, involving 548 samples across Europe, laid the groundwork for this ecotype classification. This research identified four distinct ecotypes, each with specific host preferences. ecotype I was the most widespread, infecting various livestock species. ecotype II was predominantly associated with roe deer, while ecotype III was rarer and mainly found in rodents and ticks. A fourth ecotype, infrequently observed, was found in birds and showed significant genetic divergence from the other groups. Overall, the study uncovered 97 haplotypes, revealing extensive genetic diversity within *Ap* (Jahfari et al., 2014). This molecular diversity provides insight into the epidemiological dynamics and host interactions of *Ap* across different regions and species.



Ecotype I had the broadest host range, though it notably lacked rodents or birds, marking it as a host generalist. The generalist feeding nature of *I. ricinus* nymphs and adults likely facilitates the continuous exchange of ecotype I strains in the environment. All human cases trace back to ecotype I, suggesting a high risk for zoonotic infections, though Europe has reported fewer than a few hundred human cases (Matei et al., 2019). This suggests the possibility of further subdivisions within ecotype I, given the limited human data and the diversity within the ecotype. More recently, Jaarsma et al. 2019 and Grassi et al. 2021 identified subdivisions and novel haplotypes within the four ecotypes however the same overall lineages are preserved. The ecotypes derived from *groEL* are still extensively used by researchers in Europe due to the lack of whole genome data necessary for constructing robust representations of *Ap* evolution. There is an acute need for more whole genome data to comprehensively explore strain diversity and host tropisms. In the UK, studies using the *groEL* gene have identified the presence of three of the four European ecotypes in a wide range of vertebrate species, yet no complete representations of *Ap* genomes have been generated, limiting the scope of UK-based epidemiological studies (Apa et al., 2023; Bianchessi et al., 2023; Gandy et al., 2022).

The prevalence of infection in rodents appears to be lower than in wild ruminants (Barandika et al., 2007; Silaghi et al., 2012; Stuenkel et al., 2013; Barakova et al., 2014). Phylogenetic analyses targeting the *groEL* (Jahfari et al., 2014), *msp4* (Barakova et al., 2014) and *ankA* (Majazki et al., 2013) genes reveal that rodent strains belong to a phylogenetically distinct cluster. It is therefore unlikely given these studies that rodent associated strains of *Ap* are involved in the epidemiological cycles of other mammalian hosts (Bown et al., 2009; Blaoarova et al., 2014). As previously mentioned in chapter 3 and illustrated in figure 20, at least four independent epidemiological cycles have been identified in Europe, that involve distinct hosts and vectors. Ecotype III strains have been detected in a diverse range of small vertebrate hosts including shrews, and species of rodent such as voles and mice (Bown et al., 2011). Bown et al., 2009 demonstrated that ecotype III variants of *Ap* were not transmissible by *I. ricinus* ticks further separating the lineage. Rodents are natural reservoirs for tick-borne diseases with *Ap* being no exception to this rule likely harbouring a variety of variants that are yet to be discovered (Stuenkel, 2007).

Rodents are known to exhibit high tick burdens, due to their key role in feeding *Ixodes* complex larvae, nymphs and to a lesser extent adults (Hofmeester et al., 2017; Takumi et al., 2019). Critically, rodent populations can be directly correlated with the abundance of nymph populations, and as a result the prevalence of certain tick-borne diseases (Ostfeld et al., 2001). The role of rodents in the spread of *Ap* is however not fully understood with the prevalence of the pathogen sparsely documented in Europe and very variable among species and localities (Stuen et al., 2013).

In the USA, *Ap* infection prevalence varies from 1.8% to 88.4%, depending on the study and rodent species in question (Stuen et al., 2013). It is however known that rodent associated variants of *Ap* in the USA are vastly different to those found in Europe with rodents likely serving as key reservoir hosts for *Ap*-Ha variants (Levin et al., 2002; Massung et al., 2003). Interestingly, in Asia, *Ap* has been detected in at least 16 different rodent species (Stuen et al., 2013). Multiple species of rodent such as *Rattus norvegicus* and *Apodemus agrarius* have notably high infection rates across the continent with up to 55.5% and 23.6% of individuals testing positive respectively (Stuen et al., 2013). This implies these rodent species may play a key role as reservoir hosts for rodent associated variants in Asia. *Apodemus agrarius* in particular is considered one of, if not the most important reservoir host of *Ap* in Asia (Zhan et al., 2010; Jin et al., 2012; Yang et al., 2013), however the reservoir competence of the species remains to be fully demonstrated (Zhan et al., 2008). A more recent study into rodent variants of *Ap* in Asia demonstrated that *Marmota himalayana* represents another competent reservoir host of *Ap* with an infection rate between 19.21% and 24.59% in captured marmots and marmots found dead respectively (Duan et al., 2022). Crucially, this group successfully generated a draft genome through BALB/c mice inoculation and propagation of *Ap* in cell culture (L929 cells). Whole genome comparisons revealed a diverse variant of *Ap* that may provide a reference point for future rodent associated *Ap* genomes assembled from European wildlife.

Building on the insights from Chapter 4, this final results chapter aims to utilise optimised sequencing methods to investigate the diversity and genetic composition of the European ecotypes in a global context. The Dutch National Institute for Public

Health and the Environment (RIVM) has provided the first ecotype IV and human-derived European *Ap* genomes for analysis, whilst I have attempted to generate a complete *Ap* genome derived from a common shrew identified as a representative of ecotype IV. These genomes will be instrumental in exploring *Ap* diversity and determining whether a reclassification of the species is needed to accurately describe the epidemiologically distinct clusters of strains within this highly diverse parasite.

## 5.2: Methods

### Sample Collection & Dissection

Small mammals were trapped at four sites across Kielder Forest, UK, in June 2023 using Ugglan traps baited with either hamster food (seed mix) and carrots, or carrots dipped in duck fat. Traps were checked twice daily (morning and evening) and trapped animals were removed and euthanised with isoflurane then cranial dislocation according to Home Office Schedule 1 permitted procedures. Blood samples were extracted from all individuals by cardiac puncture and transferred into 1.5ml Eppendorf tubes containing 2mg/ml of EDTA before being frozen at -20oC for storage. Carcasses were dissected and spleens removed into sterile Eppendorf 1.5ml Eppendorf tubes then frozen at -20oC for storage.

### Monarch HMW DNA Extractions

Monarch HMW DNA extractions were performed on all samples following the manufacturers guidelines for the T3060L Monarch HMW DNA Extraction from Tissue kit. DNA quality, quantity and molecular weight were assessed with three platforms, Nanodrop, Qubit and Tapestation respectively.

### *Msp2* rt-PCR

A previously described real-time PCR assay (Courtney et al., 2004) was utilised for the high sensitivity detection of *Ap* using primers: *ApMSP2f* (5'-ATGGAAGGTAGTGTTGGTTATGGTATT), *ApMSP2r* (5'-TTGGTCTTGAAGCGCTCGTA). This generated a 77bp fragment using TaqMan probe *ApMSP2p-FAM/BHQ*: (5'-TGGTGCCAGGGTTGAGCTTGAGATTG) Which is

dual labelled with FAM/BHQ. The PCR was performed with a reaction volume of 25µL using the Brilliant Quantitative PCR core reagent kit with SureStart Taq DNA polymerase in a Bio-Rad Opticon thermal cycler. Reaction conditions were as follows: primers were concentrated at 900nM each, the probe, ApMSP2p-FAM/BHQ at 125nM, with 2µL of template DNA. Cycling conditions included an initial activation of the Taq DNA polymerase for 10 minutes at 95°C, followed by 40 cycles of a 15 second denaturation at 95°C, followed by a 1-minute annealing extension step at 60°C.

### NEB Microbiome Enrichment

The CS5 Monarch HMW DNA extraction was enriched with the NEB microbiome enrichment kit for a total of eight times, with an input of 1ug in 20µL buffer of gDNA per enrichment. MBD2-Fc-bound magnetic beads were prepared according to manufacturer's instructions. Enriched microbial DNA was captured and purified using Agencourt AMPure XP beads following option A guidelines in the NEB microbiome enrichment kit manual. Purified enrichments were verified using the Qubit 3.0 fluorometer to assess quantity and the TapeStation 4150 to assess molecular weight.

### Agilent SureSelect Library Preparation

Prior to Library preparation, purified NEB enriched DNA samples were fragmented using the Bioruptor Pico (Diagenode). Enriched DNA was diluted into 50µL containing ~400ng DNA in 1.5ml Eppendorf tubes for all eight enrichments. The Bioruptor Pico was prepared according to manufacturer's instructions, cooling fresh purified water to 4°C. DNA fragmentation for fragments ranging in size from 500bp - 700bp was achieved using 2 cycles of 25 seconds ON, 30 seconds OFF ensuring samples were placed on ice prior to fragmentation for 15 minutes. Once fragmentation was complete samples were spun at 300rpm for 1 minute. Fragmentation was confirmed on the tapestation 4150 using D1000 high sensitivity screen tape according to manufacturer's guides.

The Agilent SureSelect XT HS2 DNA System protocol (Version E0, July 2022) was followed to prepare eight libraries for the CS5 samples. Notably, DNA quantity was doubled to 400ng per library and hybridisation was performed overnight to improve

capture efficiency. A total of 20 cycles was required to amplify post-capture libraries back up to suitable concentration for sequencing whilst minimising duplication.

The MiSeq V3 300 cycle kit was utilised to sequence all eight libraries. The reagent kit was thawed and prepared according to manufacturer's instructions. Library concentrations were calculated with the Qubit 3.0 fluorometer with the high sensitivity assay kit and the tapestation 4150 with D1000 screentape and reagents. Libraries were then pooled into a single equimolar library, denatured and diluted to 12.5pM prior to loading into the V3 reagent cartridge.

### PromethION P2 Solo Adaptive Sampling

Long read sequencing was performed with the PromethION P2 solo using ONT's V14 chemistry and a PromethION R10.4.1 flow cell. NEB enriched Monarch HMW DNA was sequenced using the native barcoding kit 24 V14 (SQK-NBD114.24) according to the 'ligation sequencing gDNA – native barcoding kit 24 V14' protocol available on ONT's website. The library was quantified using the Qubit 3.0 fluorometer and the Agilent tapestation 4150 using genomic screentape and reagents. The AKS2020-120P and Harris *Ap* genomes were combined into a single fasta file and specified as an enrichment target for AS. The P2 solo system was then run for 24 hours before the flow cell was washed and reloaded with excess library using the wash kit EXP-WSH004-XL following the manufacturer's guidelines. The library was then run for a further 48 hours. Basecalling was performed by Dorado using the high accuracy setting. Reads were deposited in fastq format and concatenated into a single file.

### Phylogenetics & Linkage Disequilibrium Analysis

*groEL* gene sequences were extracted from *Ap* whole genomes, GRD08, and CS5. In addition to these *groEL* fragments representatives from each European ecotype were extracted from the Jahfari et al., 2014 papers supplementary materials and chosen based on sequence length to maximise the power of the analysis.

Alignments of the *groEL* fragments was achieved in MEGA11 and sequences were trimmed appropriately. The MEGA11 maximum-likelihood analysis branch was then utilised to generate a phylogenetic tree of the alignment. Both the tree and alignment

were entered into Haploview which produced a linkage disequilibrium analysis of the haplotypes. The aesthetics of the analysis were finalised in Inkscape.

---

## 5.3: Results

A total of 13 common shrews (*Sorex araneus*) and 3 field voles (*Microtus agrestis*) were trapped and included in the study. Among these, four shrews (CS4, CS5, CS8, CS10) tested positive for *Ap* using qPCR, with Ct values recorded at 26.5, 23.5, 25.9, and 26, respectively. The shrew CS5, which exhibited the lowest Ct score of 23.5, was selected for further analysis. Using the optimised sequencing protocols described in Chapter 4, the genomic DNA from CS5 was sequenced with both the Illumina MiSeq and PromethION P2 Solo platforms. This process yielded 8,562 short reads from the Illumina MiSeq and 11 long reads from the PromethION P2 Solo. Alignment of the sequenced reads to the Harris reference genome indicated that 17.3% of the genome was covered, with only 0.2% achieving a coverage depth of 30X. The gaps in coverage were evenly distributed across the reference genome, ranging in size from approximately 100 bp to 60 kb (Figure 45). Due to the fragmentation of the genome limited conclusions could be drawn from the data. Importantly, the *groEL* gene was among the captured sequences, enabling subsequent analyses for strain identification. The *groEL* gene was extracted from the CS5 genomic data and aligned against 45 previously characterized *groEL* sequences.

This alignment included sequences from novel genomes (FG22045 and FG22047), as well as the GRD08 sequence referenced in Chapter 4. Strains FG22045 and FG22047 are complete genomic representations of two novel variants of *Ap* derived from a French *Ixodes frontalis* tick and human blood from a Slovenian patient suffering with HGA. The genomes were completed by a group at the Dutch National Institute for Public Health and the Environment (RIVM) using ONT long-read sequencing technology. Assembled genomes were kindly donated to be included in this study by Dr Hein Sprong with an ultimate goal to collaborate on the reclassification of *Ap* based on the conclusions drawn from the complete genomes of all four ecotypes. The dataset also encompassed sequences representing different ecotypes derived from the Jahfari et al., 2014 supplementary materials: KC583434

(ecotype II, Russia), KF031393 (ecotype III, Italy), and KF701461 (ecotype IV, Russia). Following trimming and alignment, a total of 23 haplotypes were identified from a 648 bp fragment of the *groEL* gene. Linkage disequilibrium analysis conducted with Haploview revealed distinct separation among the four known ecotypes. Notably, North American strains formed a subdivision within ecotype I, while the South Korean zoonotic strain KZ-A1 was classified as ecotype II. Strains CS5 and AKS2020-120P were identified as part of ecotype III, and the FG22045 strain was categorized under ecotype IV. These findings support the classification of strains within existing Eurocentric ecotypes and highlight the need for further investigation into the diversity and potential reclassification of *Ap* as more genomic data become available.

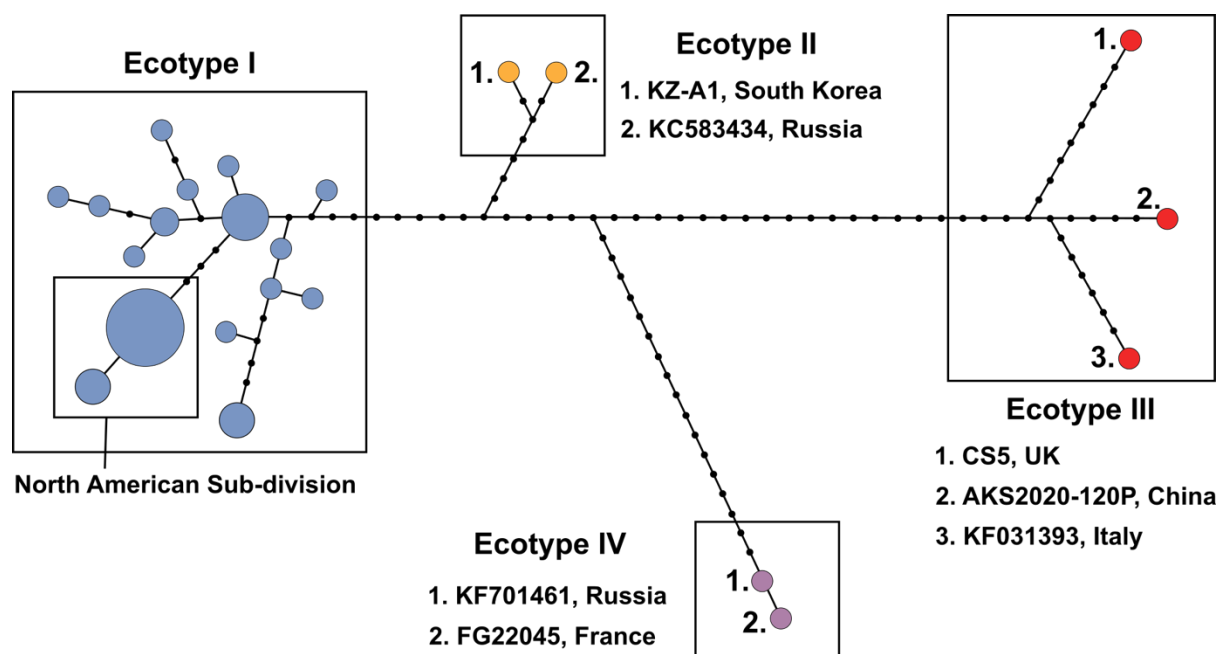
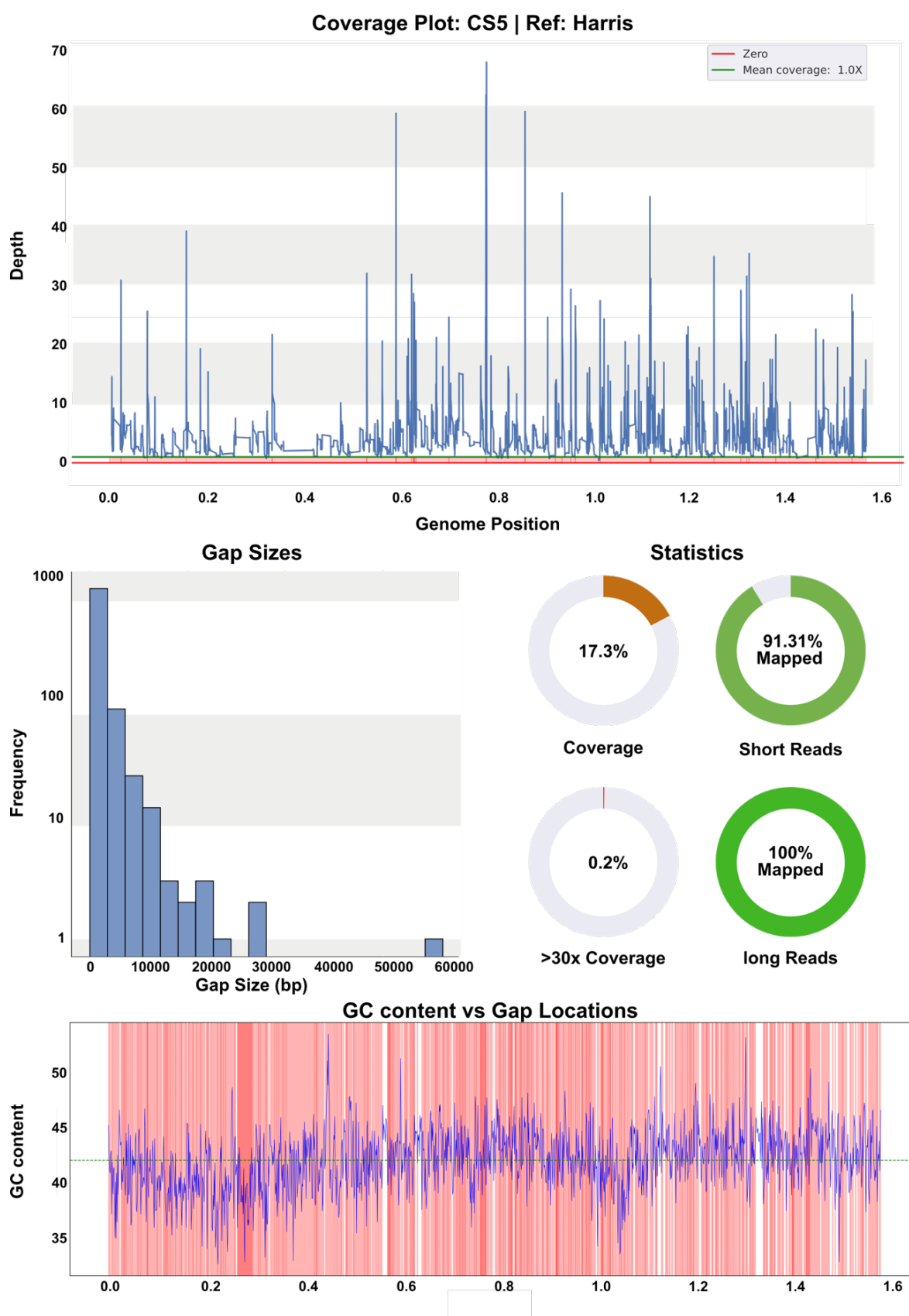


Figure 44: Linkage disequilibrium analysis of 46 *Anaplasma phagocytophilum* strains, displaying 23 haplotypes across four main ecotypes. *GroEL* fragments (648bp) aligned with Muscle integrated into MEGA11. MEGA11 was used to generate a maximum-likelihood tree bootstrapped with 100 replicates. Alignment and tree data was inputted into the Haploviewer software package.

Table 6: ANI matrix of proposed *Anaplasma phagocytophilum* ecotype representatives calculated with PyANI.

	Harris	FG22045	KZ-A1	AKS2020-120P
Harris	100	89.4	93.17	90.12
FG22045	89.4	100	89.69	92.85
KZ-A1	93.17	89.69	100	90.68
AKS2020-120P	90.12	92.85	90.68	100



**Figure 45: Coverage statistics for reads generated for a common shrew derived *Anaplasma phagocytophilum* (Ap) strain CS5 against a reference UK derived Ap strain, Harris. Top plot was**



generated in with bam2plot and represents coverage; pie charts support this data providing values for coverage and percentages of reads determined to be *Ap* group with kraken2 that were mapped onto the genome. Gap size and location analysis was generated with custom python scripts available in the GapPlotter GitHub repository.

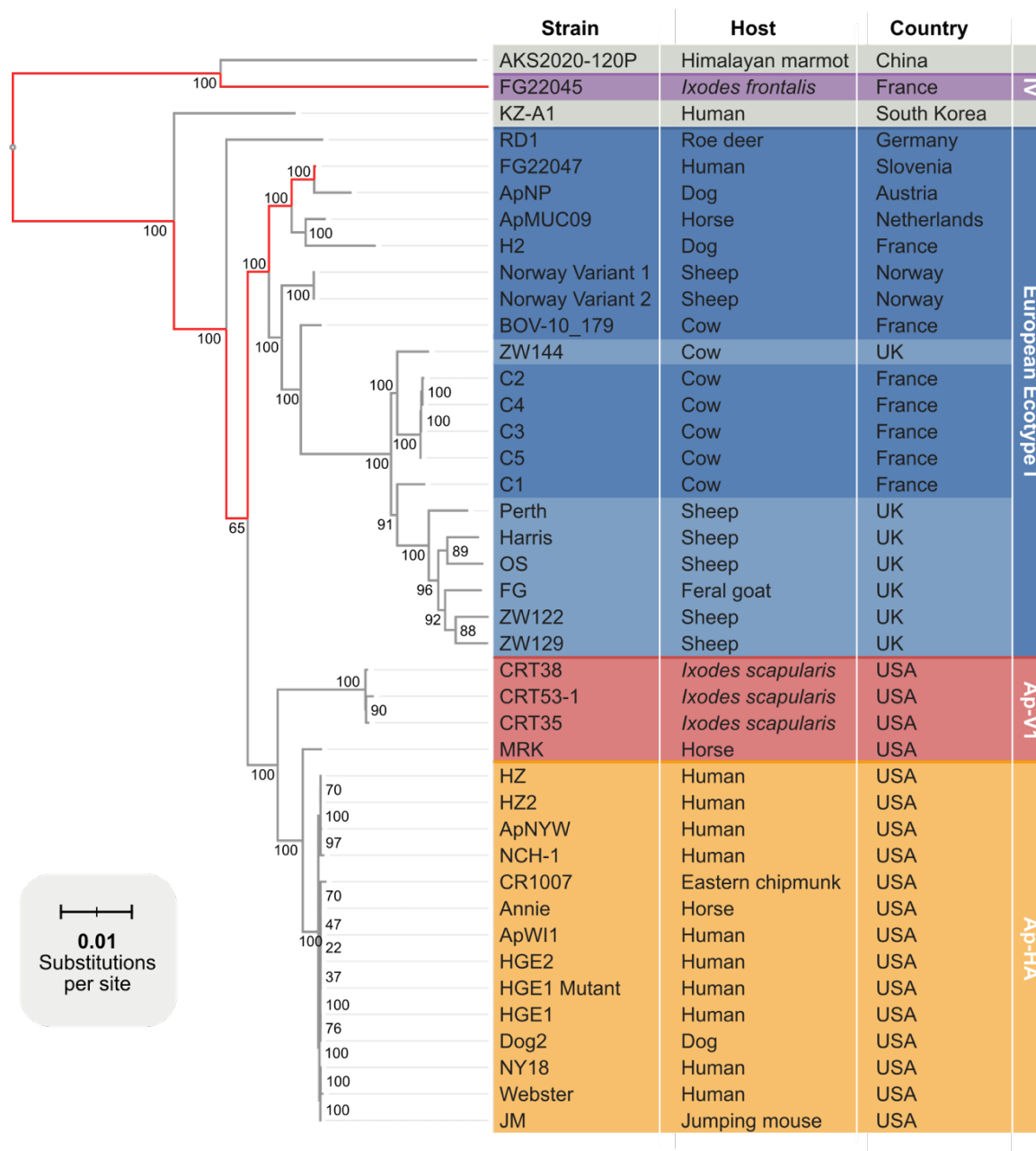


Figure 46: Maximum-likelihood phylogenetic tree generated from a concatenated alignment of amino acid sequences of 500 single copy core genes within the *Anaplasma phagocytophilum* (*Ap*) species, generated with RAxML using the DUMMY2 protein model. Scale bar represents 0.01 amino acid substitutions per site and the colours delineate known clusters/ variants. Dark blue represents strains of European origin, determined to be ecotype I. Light blue represents UK derived ecotype I strains, yellow represents the North American human active variants of *Ap* (*Ap*-HA) and red represents the North American variant 1 strains (*Ap*-V1).

A phylogenetic tree was constructed from a concatenated alignment of amino acid sequences of 500 single-copy core genes across 41 *Anaplasma phagocytophilum* genomes using RAxML with the DUMMY2 model. The tree displayed clear delineation between North American, European Ecotype I, KZ-A1 (Ecotype II-like), FG22045 (European Ecotype IV), and the rodent-associated strain AKS2020-120P (Ecotype III-like). European strain diversity was notably higher than that sequenced so far in the USA. The RD1 strain, under linkage disequilibrium analysis, appeared to be a typical Ecotype I strain, but whole-genome analysis revealed greater diversity within other single-copy core genes. The FG22047 genome, shared by RIVM, represents the first human-associated European *A. phagocytophilum* strain and provided clues into possible sub-divisions within European Ecotype I, with veterinary-associated strains clustering separately from human, dog, and horse strains.

AKS2020-120P, FG22045, and KZ-A1 strains were the most divergent, with ANI as low as 90.12%, 89.40%, and 93.17% respectively, compared to the Ecotype I UK strain Harris. A full ANI matrix between proposed ecotype representatives (Table 6). All ecotypes have ANI scores lower than the 95% species threshold but do share highly similar gene content. Genome size remains the biggest factor that can distinguish *Ap* ecotypes with the proposed ecotype III and IV strains, FG22045 and AKS2020-120P boasting sizes of 1.37Mb and 1.26Mb respectively. In comparison, KZ-A1 (ecotype II-like) and Harris (ecotype I) have undergone genome expansions, with sizes of 1.45Mb and 1.56Mb. Despite the changes in genome architecture, GC content remains stable across the ecotypes, as does gene content and redundancy.

## 5.4: Discussion

This study successfully captured and partially sequenced a novel *Ap* strain, CS5, from an infected common shrew (*Sorex araneus*) in Kielder Forest, UK. CS5 represents the first Ecotype III strain sequenced from the UK, contributing valuable genomic data to our understanding of the diversity and host associations of this zoonotic pathogen. The highly divergent nature of this ecotype from the reference genomes used for bait design (Ecotype I) limited capture efficiency. As a result, only 17.3% of the CS5 genome was successfully recovered for analysis. The difficulties encountered in sequencing CS5 highlight inherent limitations of bait capture

approaches for genomes differing substantially from the references used for bait design. Based on *groEL* linkage analysis and similarities to the complete AKS2020-120P genome derived from a Himalayan marmot, we can predict the CS5 strain to have an ANI of  $\leq 90\%$  compared to Ecotype I references such as Harris (Genome utilised for Alignment). At such levels of divergence, bait binding becomes less efficient, resulting in uneven and incomplete capture across the target genome. While bait capture remains a powerful tool for enrichment of specific genomic regions, alternative methods may be necessary for comprehensively accessing highly divergent genomes like those belonging to ecotype III. The identification of CS5 in a shrew host provides further evidence that small mammals like shrews and rodents can serve as reservoirs for certain *Ap* variants. Small mammal associated strains in the United States appear to be notably different to those encountered in Europe, we therefore cannot predict the zoonotic potential of European ecotype III variants (Aardema, 2023). The detection of an ecotype III strain aligns with previous studies reporting this ecotype primarily in rodents and insectivores across Europe (Bown et al., 2011; Silaghi et al., 2012).

In addition to the CS5 ecotype III strain, this study incorporated novel genomic data for representatives of the other three established *Ap* ecotypes through collaboration with the Dutch National Institute for Public Health and the Environment (RIVM). Notably, the FG22045 genome provides the first glimpse into the genetic makeup of the elusive ecotype IV, which has previously only been detected through genotyping studies in birds and remains poorly characterised. FG22045 was sequenced in an *Ixodes frontalis* tick and the inclusion of this genome allowed us to confidently place ecotype IV in phylogenetic context for the first time. Perhaps most significantly, the FG22047 genome represents the first European human-derived *Ap* strain sequenced from a clinical case in Slovenia. While HGA is well-documented in the United States (Telford et al., 2024), relatively few confirmed cases have been reported in Europe despite widespread presence of *Ap* in wildlife and tick populations (Bollavaram et al., 2024). Analysis of FG22047 revealed its clustering within Ecotype I, the primary zoonotic ecotype, but distinct from other European veterinary and US *Ap*-HA zoonotic variants. This suggests potential intra-ecotype subdivisions that could influence host range and pathogenicity. The availability of this first European human

isolate genome will facilitate crucial comparative studies investigating the genomic basis of human virulence and potential emergent risks of zoonotic strains.

Phylogenetic analyses based on the *groEL* gene and whole genome data supported the current delineation of *Ap* into four main ecotypes, as proposed by Jahfari et al. 2014. The *groEL* linkage disequilibrium analysis utilised representative strains from both Europe and other regions to evaluate whether the established European ecotypes accurately capture the full global diversity of *Ap*. Strains like AKS2020-120P from China and KZ-A1 from South Korea clustered with the Ecotype III and Ecotype II groups respectively. This suggests the four European ecotypes delineated by Jahfari et al. 2014 may represent the major phylogenetic divisions and ecological adaptations present across the globe. However, it is important to note that geographic distribution can still significantly impact the genetic signatures observed, even within established ecotypes. Regional populations are likely to diverge over time due to processes like genetic drift and local adaptation to regional wildlife and vectors. This effect is clearly evident in *Borrelia burgdorferi*, the causative agent of Lyme disease, which shares *I. ricinus* as its primary vector species in the UK (Barbour, 2018). While the core divide into four ecotypes appears consistent, sampling more densely across regions may reveal further sub-structuring representing geographic variation much like that observed in the more comprehensive MLST scheme by Huhn et al., 2014. This diversity does however account for less impactful differences between strains in the context of their broader ecotype associations. An additional consideration is the potential for highly divergent or novel *Ap* strains circulating in understudied regions, especially across Asia and the Pacific where a wider diversity of tick vectors have been implicated in transmission (Kuo et al., 2018). The *Ixodes ricinus* complex is considered the primary vector in Europe, while over 20 different tick species from genera like *Haemaphysalis*, *Dermacentor*, and *Rhipicephalus* have been found carrying *Ap* in parts of Asia (Rar & Golovljova 2011; Ghafar et al., 2012). These alternative vectors could harbour separate enzootic cycles involving divergent *Ap* strains that do not align cleanly with the ecotypes established based on European datasets. *Bartonella* sp. for example serve as compelling examples of how pathogens adapt to specific vectors through ecological interactions and environmental pressures (Chomel et al., 2009). Their ability to establish complex relationships with both arthropod vectors

and mammalian hosts underscores the importance of studying these dynamics in similar organisms such as *Ap*. Expanded genomic sampling, especially from other continents like Asia where diverse transmission cycles may exist, will be crucial to comprehensively describe the genetic diversity within *Ap* on a global scale. While the studies here provide a robust ecological and genomic framework, integrating data from understudied regions and vectors could reveal novel lineages or greater complexity underlying this bacterial species' diversity.

ANI values between proposed ecotype representatives were all below the 95% threshold typically utilised for bacterial species delineation, ranging from 89.40% - 93.17% as shown in table 6. This provides evidence supporting the reclassification of *Ap* into four distinct, genomically separated species. However, ecotypes I and II, although distinguishable, may in fact represent a spectrum of diversity that ultimately blurs the boundaries between these two lineages. I believe this to be the case solely for these two ecotypes as there is clear evidence that these variants regularly overlap and interact within both vector and host species (Jahfari et al., 2014; Huhn et al., 2014; Gandy et al., 2022). In contrast, ecotypes III and IV appear to be entirely distinct entities with much more substantial boundaries founded in both their vector and host species preferences (Bown et al., 2009; Jahfari et al., 2014; Huhn et al., 2014). The evolutionary trajectories of these ecotypes have likely diverged to a greater degree as their reservoir hosts and tick vectors largely separate them such that they do not interact or encounter each other in the same ecological niches. This increased isolation has driven greater genetic variation and subsequent evolution into separate but highly related tick-borne diseases. For ecotypes I and II, which overlap in vector species like *Ixodes ricinus* and can utilize a broader range of mammalian reservoir hosts, frequent opportunities for genetic exchange and introgression may occur. This could maintain a Semi-permeable barrier between these two ecotypes, allowing a spectrum of diversity to persist. Conversely, the distinct host and vector specialisations of ecotypes III and IV create more impermeable barriers, promoting the accumulation of greater genetic differences as their evolutionary trajectories remain segregated. While the four ecotypes exhibit diverse ecological adaptations, their ANI values and overall genomic similarity argue that they still represent closely related populations within the *Ap* species complex. Notably, the core genes and metabolic pathways are largely conserved (Figure 33)

despite variations in genome architecture and accessory gene repertoires across the ecotypes. This pattern suggests they may have diverged relatively recently from a common ancestor through processes like niche adaptation and genome evolutionary mechanisms like horizontal gene transfer. Therefore, reclassification into four distinct species may ultimately be warranted based on the ecological boundaries evidenced here. However, capturing the full spectrum of diversity, especially the semi-permeable barrier between ecotypes I and II, will likely require denser genomic sampling and integrating analyses of ecological interactions, host ranges, and gene flow patterns. Further research disambiguating these relationships is crucial for robustly delineating the taxonomic boundaries and developing a classification system that accurately reflects the evolutionary processes structuring diversity within this complex zoonotic pathogen.

While this study provided valuable new genomic data, identified methodological limits of bait capture and discussed the efficacy of species reclassification, several key knowledge gaps remain. The factors driving the initial formation and ecological segregation of ecotypes are still elusive. Some studies suggest links to animal migrations, such as the last glacial maximum and subsequent retreat, which drove the evolution of Ecotypes I and II as species of deer recolonised the European mainland 1500 years ago (Aardema et al., 2022). However, more research is needed to understand the broader variation across the species globally. Furthermore, denser sampling and integration of *Ap* data, especially in understudied regions of Asia and the Pacific, are required to determine the existence of additional ecotypes. Key areas for future research include the risks of emergence, host range expansion, spatial overlap of ecotypes, and the roles of genomic variations in pathogenesis and virulence in humans and domestic animals. Overall, this work contributes the first partial Ecotype III *Ap* genome sequence from the UK and highlights the remarkable diversity present within this zoonotic pathogen species. The genomic and phylogenetic analyses largely reinforce the current ecotype paradigm while also suggesting additional sub-structuring may exist, particularly in the epidemiologically important Ecotype I. Ongoing genomic surveillance and multidisciplinary studies integrating evolutionary and ecological data are warranted to fully characterise the diversity of *Ap* for improved understanding of its dynamics, risk assessment, and mitigation of public health threats.

# CHAPTER 6

## 6.0: Unravelling the Diversity & Mysteries of Blood-Borne Pathogens in the UK & Beyond – A Comprehensive Discussion.

---

### 6.1: Introduction

In the rapidly evolving field of microbiology, understanding the genetic diversity and evolutionary dynamics of pathogenic bacteria, particularly those of clinical and agricultural importance is crucial. This thesis adds to the growing body of knowledge of two blood-borne bacterial lineages, the *Bartonella* genus and *Ap* addressing key gaps in understanding through the use of specialised molecular and evolutionary analyses. The research problem centres on elucidating the genetic diversity, evolutionary processes, and epidemiological cycles of these diverse bacterial parasites, which may have substantial implications for future control strategies. Specifically, this study aims to characterise and classify three novel strains of bartonellae, termed *Bartonella bennettii* sp. nov. using hybrid sequencing techniques, contemporary bioinformatical software and biochemical and phenotypic analyses. In addition to this, I investigated the genetic diversity and epidemiological patterns of *Ap* in the UK, Europe and beyond, focusing on the characterisation of seven UK strains, the optimisation of sequencing methodologies for the capture and sequencing of *Ap* strains directly from infected tissue, and the investigation of epidemiologically separate ecotypes and their complex evolutionary relationships.

This research tested several hypotheses related to the genetic and evolutionary dynamics of these bacteria. Firstly, I hypothesised that *Bartonella bennettii* is a distinct species within the *Bartonella* genus, characterised by unique genomic features and evolutionary relationships. Secondly, I postulated that the genetic diversity of *Ap* in the UK exhibits significant regional variation, influenced by specific

ecological and evolutionary factors. Thirdly, I hypothesised that the whole genome sequencing of *Ap* strains can be achieved through the utilisation and optimisation of targeted sequencing technology. Lastly, I proposed that the formation and segregation of *Ap* ecotypes are driven by specific host-vector interactions and ecological adaptations leading to distinct evolutionary trajectories. This research is significant for several reasons; by characterising a novel species, *Bartonella bennettii*, I contribute to the growing taxonomic and phylogenetic understanding of the *Bartonella* genus, enhancing insights into the genetic exchange of virulence factors within the genus. Through the characterisation of seven UK *Ap* strains I provide the first whole genome data of the pathogen derived from UK livestock, which aids in our understanding of the genetic diversity and epidemiology of *Ap* in the UK. The employment of advanced and highly specific genomic techniques for the capture and sequencing of *Ap* strains from tissue samples investigates the limits of current technology and identifies methods that could be used for improving diagnostic and control strategies for *Ap*, a crucial step for meaningful impacts in the public health and agricultural sector. Finally, I investigated *Ap* diversity on a global scale, contributing to a broader understanding of the evolutionary trajectories of the species, identifying potential species boundaries across the globally present ecotypes. By addressing these research aims and questions, this thesis not only fills critical gaps in current knowledge but also sets the stage for future research in microbial genomics and epidemiology, particularly the study of *Ap*.

---

## 6.2: Key Findings

Genomic and phylogenetic analysis establish *B. bennettii* as a distinct L3 species of *Bartonella*. Whole genome sequencing identified a sufficiently divergent genome with high structural similarity to other L3 species such as *B. rochalimae* and *B. clarridgeiae*. Phylogenetic trees consistently place *B. bennettii* in L3 with high bootstrap support, indicating strong confidence in the evolutionary relationships depicted. The phylogenetic placement of *B. bennettii* underscores its novelty and provides a framework for further comparative studies within the genus. *B. bennettii* possessed all the characteristic virulence factors associated with L3 which on closer inspection could be modelled to *B. clarridgeiae* and more specifically the partially characterised strain AR15-3. AR15-3 was determined to be a divergent (potential



subspecies) of *B. bennettii* due to high overall similarity (ANI 95%) with the three field vole derived strains, C271, J117 and D105. Through whole genome synteny comparisons between the strains, a unique genomic region characteristic of a chromosomally integrated plasmid was identified in the strains C271, J117 and D105, but absent in AR15-3. Due to the absence in AR15-3 we could determine the exact size of the insert to be 33kb. This integrated plasmid contained a *vbh* T4SS that may provide a unique evolutionary advantage in these strains of *B. bennettii*. Additionally, pangenomic profiling revealed significant differences in gene content between *B. bennettii* and other species of *Bartonella*. These differences include variations in genes related to metabolism, and immune evasion. Such variations suggest that *B. bennettii* has evolved unique strategies to adapt to its specific host environment, which may differ from those of other more divergent *Bartonella* species. Through the characterisation of *B. bennettii* and comparative analysis with the *Bartonella* genus we provide compelling evidence for its classification as a novel species. The unique genomic features, distinct phylogenetic position, and specific adaptations highlight the importance of this novel species and more broadly the scientific value in characterising bacterial strains and species.

Through the utilisation of tick cell cultures, chapter 3 provides the first whole genome representations of *Ap* strains collected in the UK. The successful sequencing of these strains provided critical insights into the genetic variability of *Ap* across different regions within the UK and more broadly within European ecotype 1. The data revealed significant regional variation, with genetic diversity correlating with geographical location and host species, however more information is required to fully elucidate these evolutionary relationships. This work demonstrated the effectiveness of tick cell cultures for studying and sequencing *Ap* strains, facilitating a deeper understanding of their epidemiology and evolutionary dynamics. Crucially the information generated in this chapter was instrumental in the designing of the Agilent SureSelect baits, which were used to enhance representation of *Ap* reads in metagenomic sequencing datasets in chapter 4.

Chapter 4 explored the efficacy of sequencing *Ap* strains directly from tissue samples. Through collaboration with a local Cumbrian abattoir and forestry England a total of 80 sheep spleen, 50 roe deer spleens and 23 red deer spleen were

collected and screened for *Ap* finding infection rates of 10%, 88% and 52.2% respectively. An evaluation of adaptive sampling (ONT), bait capture (Agilent), methylation depletion (NEB) and differential lysis (Molzym) was performed to find the optimal methodology for sequencing the elusive intracellular parasite. Benchmark sequencing runs found that only a fraction of reads were identified as *Ap* (<0.0014%) when sequencing unenriched HMW DNA due to poor pathogen: host ratios. Adaptive sampling initially proved successful but failed to consistently improve the representation of *Ap* in datasets. Interestingly, similarity to the reference genome appeared to be the most important factor in determining success with Ct failing to correlate with *Ap* read representation. On the other hand, bait capture using the Agilent SureSelect platform performed very well throughout experimentation. Ct correlated well with *Ap* read representation as initially predicted. Evaluations of the differential lysis kit from Molzym and the microbiome enrichment kit from NEB identified the incompatibility of the kits. The differential lysis kit failed to perform as hoped whilst the NEB microbiome enrichment kit when combined with HMW DNA extractions and bait capture performed the best. This optimal approach achieved a share of 65.74% *Ap* reads in GRD08, representing an enrichment of 49454x.

Genome coverage was quantified for the three roe deer strains sequenced. Coverage correlated with Ct values, achieving at least 1x coverage for 69.8% of GRD08 (Ct 19.0), 58.1% of GRD09 (Ct 20.5), and 54.7% of GRD17 (Ct 21.2). Coverage plateaued for all samples, particularly GRD08 which underwent more sequencing indicating that certain portions of the genome could not be captured by the baits. Sequencing depth did however increase uniformly with more reads resulting in a greater % of the genome with at least 30x coverage. The gaps in the alignment were investigated using custom python scripts written as part of the GapPlotter package available on GitHub. Perhaps most importantly, gap analysis of the three Roe Deer strain datasets indicated that read length was the primary limiting factor in the capture of whole genomes, as gaps in the alignment were largely less than a few kb, reaching a maximum of just over 7500bp in GRD17. Ultimately, adaptation of this methodology to nanopore systems would likely facilitate the complete capture of *Ap* genomes if an average read length of 4kb could be achieved.

Chapter 5 critically evaluates the effectiveness of the bait capture technology, explores global *Ap* diversity and identifies species boundaries between the epidemiologically separate ecotypes. Whilst the bait capture technology proved useful for the capture of Roe Deer derived strains that identify with European ecotype 1, the technology failed to perform with the same efficiency when targeting ecotype III strain, CS5, derived from a common shrew. Reduced capture efficiency was primarily due to the high genetic variation of ecotype III strains (<90% ANI with ecotype I). Despite these limitations the technology yielded 17.3% of the genome.

### 6.3: Contextualising Research Findings

The characterisation of *B. bennettii*, along with the first chromosomal representation of *B. hiexiaziensis*, provides valuable insights into the *Bartonella* genus. Comparative genomic analysis of 48 *Bartonella* genomes, including 32 validly published and 15 partially characterised species, revealed important patterns in the evolution of virulence factors across the genus. The presence of the *Bartonella* gene transfer agent (BaGTA) in all analysed genomes suggests its critical role in the success of *Bartonella* species, likely due to its involvement in host interactions (Berglund et al., 2009; Guy et al., 2013). The mutual exclusivity of flagella and the trw Type IV Secretion System (T4SS) across the genus, with trw T4SS present only in Lineage 4 (L4) species, implies that the acquisition of trw T4SS was a key driver in the adaptive radiation and success of L4 *Bartonella* (Deng et al., 2010; Deng et al., 2012). The virB/D4 T4SS, present in L1, L3, and L4 *Bartonella* species, appears to have been acquired multiple times throughout the genus's evolution. Its role in translocating effector proteins (Beps) into host cells for immune evasion highlights its importance as a virulence factor (Engel et al., 2011; Wagner & Dehio, 2019). The presence of functional Beps in all genomes with the virB/D4 locus further supports the hypothesis of multiple independent acquisitions. The discovery of the vbh/*TraG* T4SS in *B. bennettii*'s chromosome is particularly intriguing, as this system is present in only 9 of the 48 analysed genomes. Its potential role as an interbacterial conjugation system (Harms et al., 2015; Harms et al., 2017) and its absence in the closely related AR15-3 strain suggest recent acquisition and potential ongoing evolution within the species. The comparative analysis also revealed that rodent-associated *Bartonella* strains, including *B. bennettii*, appear to undergo more frequent recombination

events than other *Bartonella* species (Berglund et al., 2009; Paziewska et al., 2011). This higher recombination rate suggests a broader host range for rodent-adapted species and may contribute to their genetic diversity and adaptability. The pangenome analysis of the *Bartonella* genus, encompassing over 84,000 genes across 9,356 orthologous gene clusters, provides a comprehensive view of the genetic diversity within the genus. The relatively small singleton genome of *B. bennettii*, compared to species like *B. tamiae*, suggests that while it is a novel species, it shares many genetic elements with other members of the genus.

The comprehensive characterisation of *Bartonella bennettii* served as an invaluable learning experience, equipping me with a robust skill set in microbial genomics and comparative analysis. This foundation proved essential when progressing to the study of more complex microbial pathogens, particularly *Ap*. The techniques employed in the *B. bennettii* study, such as whole genome sequencing, de novo assembly, and comparative genomics, were directly applicable to the analysis of *Ap*. However, the latter presented additional challenges due to its obligate intracellular lifestyle and the complexities of its genome. The experience gained from assembling and annotating the *B. bennettii* genome using hybrid assembly methods was particularly useful when dealing with the more fragmented and repetitive genome of *Ap*. Furthermore, the comparative genomic approaches used to investigate virulence factors in *Bartonella* species provided a framework for exploring similar elements in *Ap*. The skills developed in identifying and analysing secretion systems, effector proteins, and other virulence-associated genes in *B. bennettii* were instrumental in unravelling the complex host-pathogen interactions of *Ap*. In essence, the study of *B. bennettii* served as a crucial stepping stone, allowing for the development and refinement of bioinformatic and genomic analysis skills that were subsequently applied to the more challenging investigation of *Ap*.

The genome of *Ap* has been a subject of significant interest due to its role as an emerging zoonotic pathogen and its impact on livestock. Despite its importance, relatively few complete *Ap* genomes had been sequenced prior to this study. Only 33 complete genomes were available across both the NCBI and ezbiocloud databases (Battilani et al., 2017). This limited genomic data has hindered our understanding of *Ap*'s genetic diversity, host adaptation mechanisms, and virulence factors. Our study

makes a substantial contribution by providing the first seven complete *Ap* genomes from the United Kingdom. These genomes, derived from strains isolated from Scottish sheep (Harris, Old Sourhope, Perth), Welsh sheep (ZW122), English sheep (ZW129), English cattle (ZW144), and Scottish feral goats (FG), represent a significant expansion of the available genomic data for *Ap* (Foster & Cameron, 1970; Scott & Horsburgh, 1983; Woldehiwet et al., 2002; Woldehiwet and Horrocks, 2005). The importance of these genomes lies not only in their geographical origin but also in their potential to reveal UK-specific adaptations or genetic features. The sequenced UK *Ap* genomes were found to be typical of ecotype I, which aligns with previous research identifying this as the most common and epidemiologically relevant ecotype for livestock farming (Jahfari et al., 2014). The genome sizes ranged from 1.52 Mb to 1.61 Mb, with GC content varying from 41.29% to 41.84%, consistent with the high GC content reported for *Ap* compared to other Rickettsiales (Battilani et al., 2017). Our comparative genomics approach built upon previous studies by providing a more comprehensive view of *Ap* genetic diversity. The phylogenetic analysis based on 500 single-copy core genes revealed clear delineations between Asian, European, and North American strains. Notably, our analysis showed that European strains, including our UK isolates, exhibited higher genetic diversity compared to the North American strains. This finding supports and extends previous observations of genetic diversity within *Ap* strains (Jahfari et al., 2014; Battilani et al., 2017). Our pangenome analysis identified 1,660 homologous gene clusters containing 9,910 total genes across the examined *Ap* strains, providing a more comprehensive view of the *Ap* genome than previous studies. The high proportion (77.49%) of genes annotated as hypothetical proteins underscores the need for further functional characterisation, a challenge also noted in earlier genomic studies (Battilani et al., 2017).

ANI analysis revealed high genetic relatedness across most *Ap* genomes, consistent with previous findings. However, our study went further by identifying potential novel species classification boundaries when comparing certain divergent strains (e.g., AKS2020-120P and KZ-A1) with the well-represented western variants, suggesting a greater complexity in *Ap* taxonomy than previously recognised. By contributing these seven UK *Ap* genomes and performing comparative genomic analyses, our study significantly expands upon previous research. While earlier studies like Jahfari et al. 2014 established the ecotype classification, our work provides a more nuanced

understanding of genetic diversity within ecotype I, particularly in the UK. Similarly, while Battilani et al. 2017 noted genetic diversity in *Ap*, our study quantifies and contextualises this diversity on a broader scale, particularly highlighting the higher diversity in European strains compared to North American ones. Our findings also build upon the work of Bown et al. 2007, who initially investigated the ZW122, ZW129, and ZW144 strains using VNTR loci. By providing complete genome sequences for these strains, we offer a much more comprehensive genetic characterisation, allowing for more detailed comparisons and analyses. Furthermore, our study complements and extends the research of Woldehiwet et al. 2002 and Woldehiwet and Horrocks 2005, who established these *Ap* strains in tick cell cultures. By sequencing the genomes of these cultured strains, we bridge the gap between *in vitro* studies and genomic analyses, providing a valuable resource for future investigations into host-pathogen interactions and virulence mechanisms.

Studying the epidemiology of diseases such as *Ap*, requires a significant understanding of the transmission dynamics between both domestic and wild populations of susceptible animals (Gandy et al., 2022). PCR identification of *Ap*, either through highly sensitive real-time assays targeting the *msp2* genes (Courtney et al., 2004) or standard assays targeting the *groEL* gene (Jahfari et al., 2014) have formed the backbone of *Ap* epidemiological research. However, by today's standards, PCR based studies, although a critical step, fail to provide the volume of data required to accurately depict the evolutionary and epidemiological relationships between species and strains. High-throughput sequencing technologies such as those utilised to sequence the seven strains of *Ap* in chapter 3 provide orders of magnitude more information that can delineate strains more accurately with greater confidence. For example, although *groEL* based studies have identified the presence of at least four distinct European ecotypes circulating across Europe, without more information, these ecotypes remain a hypothesis yet to be proven with whole genome sequencing technology. As mentioned throughout this thesis, the whole genome sequencing of *Ap* presents a number of challenges that are difficult to overcome. The obligate intracellular nature of the pathogen almost entirely prevents its isolation and cultivation, with the exception of specialised tick-cell cultures (Woldehiwet et al. 2002). This substantially inhibits the number and types of samples that can be sequenced using traditional methodologies described in chapter 3. Recent

advancements in DNA-based technology have led to the rise of highly specific targeted sequencing technologies ranging from microbiome enrichment kits to in-silico live selection of DNA strands. One of the most exciting prospects from this relatively new set of technologies is hybridisation-capture, which can be used to enrich the proportion of subject DNA in complex metagenomic samples. Perhaps most importantly, this technology has been demonstrated for a number of applications, from sequencing low population microbes (Depledge et al., 2011) to assessing DNA damage analysis of ancient marine eukaryotes (Armbrecht et al., 2021), all whilst reducing sequencing costs through increases in efficiency (Gaudin & Desnues, 2018). The efficacy of this technology was demonstrated in chapter 4, identifying its ability to effectively capture ecotype I strains of *Ap*. We did however identify some shortfalls of the technology, for example, the capture process needed to be repeated for the same sample and pooled afterwards for optimal coverage of the genome. In addition to this, capture efficiency could be correlated with Ct of the samples with high CTs producing poor results. The final limitation of this methodology was identified in chapter 5, which indicated that the baits need to be highly homogeneous with the target strands in the sample for efficient capture. Despite these limitations, the capture of *Ap* strains from the roe deer samples provides promise for the technology especially if increased read lengths can be achieved.

## 6.4: Future Directions

The factors driving the initial formation and ecological segregation of the *Ap* ecotypes remains elusive. More research is needed to understand both localised and broader variation across the species. This requires not only denser sampling across the globe but more complete datasets. Efficiently studying *Ap* requires methodologies that bypass the expensive, time-consuming, and highly specialised cultivation steps, which in all cases may not support diverse examples of the pathogen. Therefore, it is imperative that we continue to develop and refine methodologies that can target and sequence *Ap* directly from blood and tissue samples. Optimising current techniques and exploring new technologies will be vital to achieve better results and further advance our understating of this complex highly diverse pathogen.

One of the primary limiting factors of our current enrichment methodology is the shortfalls of short read sequencing and the relative failure of AS on ONT systems to provide sufficient enrichments. To eliminate these issues, it is possible to work with Agilent technologies to develop baits compatible with the latest V14 ONT chemistry and increase read length. As illustrated by the gap analysis in chapter 4, it is theoretically possible to sequencing these three *Ap* genomes in samples GRD08, GRD09 and GRD17 if read length is increased to 5kb. Increased read lengths should bridge across regions that we have failed to capture providing more complete coverage and subsequently aiding in both mapping onto a reference genome and future hybrid assembly of these strains.

Perhaps more exciting is the application of ATAC sequencing (assay for transposase-accessible chromatin using sequencing), a technique used to study the accessibility of chromatin in the genome. It provides insights into which regions of DNA are open and accessible for transcription and other regulatory processes, which is crucial for understanding gene expression and regulation. The method involves the following steps: transposase insertion, library preparation, sequencing and analysis. ATAC-seq relies on a hyperactive transposase enzyme (Tn5) which simultaneously cuts accessible regions of DNA and inserts sequencing adapters. These accessible regions are typically those that are not tightly bound by histones or other chromatin proteins, making them more likely to be involved in regulatory activities. Neutrophils have highly condensed chromatin, limiting their accessible regions compared to other cell types due to their specialised role in the immune system, which requires limited transcriptional activity. This condensed state largely inhibits the ability of Tn5 to cut regions of DNA and insert sequencing adapters, in theory inhibiting the sequencing of host DNA. Tn5 will however cut and insert adapters into bacterial DNA within the neutrophil, provided the DNA is accessible. Therefore, collection of infected blood, cell sorting and ATAC-seq may provide a novel methodology we could exploit in future experiments to generate complete genomes of *Ap* strains.

One of the major limitations of this study was the inability to generate accurate annotations for the majority of genes predicted in the pangenome of *Ap*. Over 77% of genes were annotated as hypothetical proteins, significantly impacting the



conclusions that could be made when examining gene presence and absence. This alerts to a broader need for more powerful tools for predicting protein functions. Artificial intelligence (AI) is increasingly being used to predict and generate functions for genes. AI, particularly machine learning and deep learning techniques can analyse vast amounts of genomic data to identify patterns and make functional predictions.

## 6.5: Conclusion

This comprehensive study has significantly advanced our understanding of blood-borne bacterial pathogens. Through rigorous genomic and phylogenetic analyses, we have established *B. bennettii* as a distinct L3 species within the *Bartonella* genus, highlighting its unique genomic features and evolutionary relationships. The characterisation of *B. bennettii* not only expands our taxonomic knowledge but also provides crucial insights into the genetic exchange of virulence factors within the genus (Engel et al., 2011; Wagner & Dehio, 2019).

Our investigation of *Ap* has yielded the first whole genome representations of UK strains, offering critical insights into the genetic variability across different regions and within European ecotype I. The successful application of tick cell cultures for studying and sequencing *Ap* strains has facilitated a deeper understanding of their epidemiology and evolutionary dynamics (Woldehiwet et al., 2002; Woldehiwet and Horrocks, 2005). Moreover, the optimisation of sequencing methodologies, particularly the use of bait capture technology, has demonstrated significant potential for capturing and sequencing *Ap* strains directly from infected tissue. This advancement represents a crucial step towards improving diagnostic and control strategies for *Ap*, with substantial implications for public health and the agricultural sector. The global diversity analysis of *Ap* has contributed to a broader understanding of its evolutionary trajectories, identifying potential species boundaries across globally present ecotypes. Our findings have revealed higher genetic diversity in European strains compared to North American ones, extending previous observations and providing a more nuanced view of *Ap* genetic diversity (Jahfari et al., 2014; Battilani et al., 2017). However, this research has also highlighted significant challenges and areas for future investigation. The high proportion of hypothetical proteins identified in the *Ap* pangenome underscores the need for

further functional characterisation. Additionally, the limitations of current sequencing technologies in capturing complete *Ap* genomes from metagenomic samples emphasise the necessity for continued methodological refinement.

Future directions should focus on denser global sampling and the development of more efficient targeted sequencing technologies to bypass the need for cultivation. The potential application of ATAC sequencing and the integration of artificial intelligence for improved gene function prediction represent exciting avenues for future research. These advancements will be crucial in unravelling the complex epidemiology and evolution of *Ap*, potentially leading to more effective control strategies and a deeper understanding of host-pathogen interactions. In conclusion, this study has made significant contributions to the field of microbial genomics and epidemiology, particularly in the context of blood-borne pathogens. By addressing critical gaps in current knowledge and setting the stage for future research, this work provides a solid foundation for ongoing efforts to combat these important bacterial parasites and mitigate their impact on human and animal health.

# REFERENCES

- Aardema, M. L. (2023). Genomic analyses indicate the North American *Ap*-ha variant of the tick-vectored bacterium *Anaplasma phagocytophilum* was introduced from Europe. *Parasites & Vectors*, 16(1), 301.
- Adams, P. S., Bintzler, D., Bodi, K., Dewar, K., Grove, D. S., Kieleczawa, J., ... & Perera, A. G. (2010). Comparison of commercially available target enrichment methods for next generation sequencing with Illumina platform. *Journal of Biomolecular Techniques: JBT*, 21(3 Suppl), S17.
- Al-Khedery, B., Lundgren, A. M., Stuen, S., Granquist, E. G., Munderloh, U. G., Nelson, C. M., ... & Barbet, A. F. (2012). Structure of the type IV secretion system in different strains of *Anaplasma phagocytophilum*. *BMC genomics*, 13, 1-15.
- Al-Rofaai, A., & Bell-Sakyi, L. (2020). Tick cell lines in research on tick control. *Frontiers in physiology*, 11, 152.
- Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1), 61-65.
- Aly, K. A., & Baron, C. (2007). The VirB5 protein localizes to the T-pilus tips in *Agrobacterium tumefaciens*. *Microbiology*, 153(11), 3766-3775.
- Anderson, B., Sims, K., Regnery, R., Robinson, L., Schmidt, M. J., Goral, S., ... & Edwards, K. (1994). Detection of *Rochalimaea henselae* DNA in specimens from cat scratch disease patients by PCR. *Journal of Clinical Microbiology*, 32(4), 942-948.
- Apaa, T. T., McFadzean, H., Gandy, S., Hansford, K., Medlock, J., & Johnson, N. (2023). *Anaplasma phagocytophilum* Ecotype Analysis in Cattle from Great Britain. *Pathogens*, 12(8), 1029.
- Ambrecht, L., Hallegraeff, G., Bolch, C. J. S., Woodward, C., & Cooper, A. (2021). Hybridisation capture allows DNA damage analysis of ancient marine eukaryotes. *Scientific Reports*, 11(1), 3220.
- Atif, F. A. (2015). *Anaplasma marginale* and *Anaplasma phagocytophilum*: Rickettsiales pathogens of veterinary and public health significance. *Parasitology research*, 114, 3941-3957.
- Bai, Y., Malania, L., Alvarez Castillo, D., Moran, D., Boonmar, S., Chanlun, A., ... & Kosoy, M. (2013). Global distribution of *Bartonella* infections in domestic bovine and characterization of *Bartonella bovis* strains using multi-locus sequence typing. *PLoS One*, 8(11), e80894.
- Bai, Y., Rizzo, M. F., Alvarez, D., Moran, D., Peruski, L. F., & Kosoy, M. (2015). Coexistence of *Bartonella henselae* and *B. clarridgeiae* in populations of cats and their fleas in Guatemala. *Journal of Vector Ecology*, 40(2), 327-332.
- Barakova, I., Derdakova, M., Carpi, G., Rosso, F., Collini, M., Tagliapietra, V., ... & Rizzoli, A. (2014). Genetic and ecologic variability among *Anaplasma phagocytophilum* strains, northern Italy. *Emerging infectious diseases*, 20(6), 1082.

- Barandika, J. F., Hurtado, A., García-Esteban, C., Gil, H., Escudero, R., Barral, M., ... & García-Pérez, A. L. (2007). Tick-borne zoonotic bacteria in wild and domestic small mammals in northern Spain. *Applied and environmental microbiology*, 73(19), 6166-6171.
- Barbet, A. F., Agnes, J. T., Moreland, A. L., Lundgren, A. M., Alleman, A. R., Noh, S. M., ... & Palmer, G. H. (2005). Identification of functional promoters in the *msp2* expression loci of *Anaplasma marginale* and *Anaplasma phagocytophilum*. *Gene*, 353(1), 89-97.
- Barbet, A. F., Al-Khedery, B., Stuen, S., Granquist, E. G., Felsheim, R. F., & Munderloh, U. G. (2013). An emerging tick-borne disease of humans is caused by a subset of strains with conserved genome structure. *Pathogens*, 2(3), 544-555.
- Barbet, A. F., Lundgren, A. M., Alleman, A. R., Stuen, S., Bjöersdorff, A., Brown, R. N., ... & Foley, J. E. (2006). Structure of the expression site reveals global diversity in *MSP2* (P44) variants in *Anaplasma phagocytophilum*. *Infection and immunity*, 74(11), 6429-6437.
- Barbet, A. F., Meeus, P. F. M., Belanger, M., Bowie, M. V., Yi, J., Lundgren, A. M., ... & Jauron, S. D. (2003). Expression of multiple outer membrane protein sequence variants from a single genomic locus of *Anaplasma phagocytophilum*. *Infection and immunity*, 71(4), 1706-1718.
- Barbour, A. G. (2018). Biology of the Lyme disease agents: a selective survey of clinical and epidemiologic relevance. In *J Halperin: Lyme Disease: An evidence-Based Approach. 2nd ed* (pp. 29-44). CAB International.
- Bárdy, P., Füzik, T., Hrebík, D., Pantůček, R., Thomas Beatty, J., & Plevka, P. (2020). Structure and mechanism of DNA delivery of a gene transfer agent. *Nature Communications*, 11(1), 3034.
- Battilani, M., De Arcangeli, S., Balboni, A., & Dondi, F. (2017). Genetic diversity and molecular epidemiology of *Anaplasma*. *Infection, Genetics and Evolution*, 49, 195-211.
- Bell-Sakyi, L. (1991). Continuous cell lines from the tick *Hyalomma anatolicum anatolicum*. *The Journal of parasitology*, 1006-1008.
- Bell-Sakyi, L., Darby, A., Baylis, M., & Makepeace, B. L. (2018). The Tick Cell Biobank: A global resource for in vitro research on ticks, other arthropods and the pathogens they transmit. *Ticks and tick-borne diseases*, 9(5), 1364-1371.
- Benson, L. A., Kar, S. I. D. D. H. A. R. T. H. A., McLaughlin, G., & Ihler, G. M. (1986). Entry of *Bartonella bacilliformis* into erythrocytes. *Infection and immunity*, 54(2), 347-353.
- Bereswill, S., Hinkelmann, S., Kist, M., & Sander, A. (1999). Molecular analysis of riboflavin synthesis genes in *Bartonella henselae* and use of the *ribC* gene for differentiation of *Bartonella* species by PCR. *Journal of clinical microbiology*, 37(10), 3159-3166.
- Berglund, E. C., Frank, A. C., Calteau, A., Vinnere Pettersson, O., Granberg, F., Eriksson, A. S., ... & Andersson, S. G. (2009). Run-off replication of host-adaptability genes is associated with gene transfer agents in the genome of mouse-infecting *Bartonella grahamii*. *PLoS genetics*, 5(7), e1000546.

Berglund, E. C., Granberg, F., Zhoupeng, X., Ellegaard, K., Kosoy, M. Y., Birtles, R., & Andersson, S. G. (2009). Diversification by recombination in *Bartonella grahamii* from wild rodents in Asia contrasts with a clonal population structure in Northern Europe and America.

Bhatty, M., Gomez, J. A. L., & Christie, P. J. (2013). The expanding bacterial type IV secretion lexicon. *Research in microbiology*, 164(6), 620-639.

Bianchessi, L., Rocchi, M. S., Maley, M., Allen, K., Ballingall, K., & Turin, L. (2023). Presence of *Anaplasma phagocytophilum* Ecotype I in UK Ruminants and Associated Zoonotic Risk. *Pathogens*, 12(2), 216.

Billeter, S. A., Levy, M. G., Chomel, B. B., & Breitschwerdt, E. B. (2008). Vector transmission of *Bartonella* species with emphasis on the potential for tick transmission. *Medical and veterinary entomology*, 22(1), 1-15.

Birtles, R. J. (2005). Bartonellae as elegant hemotropic parasites. *Annals of the New York Academy of Sciences*, 1063(1), 270-279.

Birtles, R. J., & Raoult, D. (1996). Comparison of partial citrate synthase gene (gltA) sequences for phylogenetic analysis of *Bartonella* species. *International Journal of Systematic and Evolutionary Microbiology*, 46(4), 891-897.

Birtles, R. J., Fry, N. K., Ventosilla, P., Cáceres, A. G., Sánchez, E., Vizcarra, H., & Raoult, D. (2002). Identification of *Bartonella bacilliformis* genotypes and their relevance to epidemiological investigations of human bartonellosis. *Journal of Clinical Microbiology*, 40(10), 3606-3612.

Birtles, R. J., Harrison, T. G., & Molyneux, D. H. (1994). Grahamella in small woodland mammals in the UK: isolation, prevalence and host specificity. *Annals of Tropical Medicine & Parasitology*, 88(3), 317-327.

Birtles, R. J., Harrison, T. G., Saunders, N. A., & Molyneux, D. H. (1995). Proposals To Unify the Genera Grahamella and *Bartonella*, with Descriptions of *Bartonella talpae* comb. nov., *Bartonella peromysci* comb. nov., and Three New Species, *Bartonella grahamii* sp. nov., *Bartonella taylorii* sp. nov., and *Bartonella doshiae* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 45(1), 1-8.

Blaňarová, L., Stanko, M., Carpi, G., Miklisová, D., Víchová, B., Mošanský, L., ... & Derdáková, M. (2014). Distinct *Anaplasma phagocytophilum* genotypes associated with *Ixodes trianguliceps* ticks and rodents in Central Europe. *Ticks and Tick-borne Diseases*, 5(6), 928-938.

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., ... & Zianni, M. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. *Journal of biomolecular techniques: JBT*, 24(2), 73.

Bollavaram, V. K., Mittal, A., Bucaloiu, I. D., Long, J. L., & Jha, U. (2024). Double Trouble: Atypical Presentation of Human Granulocytic Anaplasmosis and Lyme Disease Co-infection. In *B43. "WE PROBABLY NEED AN ID CONSULT"-ATYPICAL INFECTIONS IN THE ICU* (pp. A3597-A3597). American Thoracic Society.

Bouhsira, E., Franc, M., Boulouis, H. J., Jacquet, P., Raymond-Letron, I., & Liénard, E. (2013). Assessment of persistence of *Bartonella henselae* in *Ctenocephalides felis*. *Applied and environmental microbiology*, 79(23), 7439-7444.

Boulouis, H. J., Barrat, F., Bermond, D., Bernex, F., Thibault, D., Heller, R., ... & Chomel, B. B. (2001). Kinetics of *Bartonella birtlesii* infection in experimentally infected mice and pathogenic effect on reproductive functions. *Infection and immunity*, 69(9), 5313-5317.

Bown, K. J., Begon, M., Bennett, M., Birtles, R. J., Burthe, S., Lambin, X., ... & Ogden, N. H. (2006). Sympatric *Ixodes trianguliceps* and *Ixodes ricinus* ticks feeding on field voles (*Microtus agrestis*): potential for increased risk of *Anaplasma phagocytophilum* in the United Kingdom?. *Vector-Borne & Zoonotic Diseases*, 6(4), 404-410.

Bown, K. J., Lambin, X., Ogden, N. H., Begon, M., Telford, G., Woldehiwet, Z., & Birtles, R. J. (2009). Delineating *Anaplasma phagocytophilum* ecotypes in coexisting, discrete enzootic cycles. *Emerging infectious diseases*, 15(12), 1948.

Bown, K. J., Lambin, X., Telford, G. R., Ogden, N. H., Telfer, S., Woldehiwet, Z., & Birtles, R. J. (2008). Relative importance of *Ixodes ricinus* and *Ixodes trianguliceps* as vectors for *Anaplasma phagocytophilum* and *Babesia microti* in field vole (*Microtus agrestis*) populations. *Applied and environmental microbiology*, 74(23), 7118-7125.

Bown, K. J., Lambin, X., Telford, G., Heyder-Bruckner, D., Ogden, N. H., & Birtles, R. J. (2011). The common shrew (*Sorex araneus*): a neglected host of tick-borne infections?. *Vector-borne and zoonotic diseases*, 11(7), 947-953.

Bradbury, C. A., & Lappin, M. R. (2010). Evaluation of topical application of 10% imidacloprid–1% moxidectin to prevent *Bartonella henselae* transmission from cat fleas. *Journal of the American Veterinary Medical Association*, 236(8), 869-873.

Brand, A., Brand, H., & Schulte in den Bäumen, T. (2008). The impact of genetics and genomics on public health. *European Journal of Human Genetics*, 16(1), 5-13.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., ... & Xia, F. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*, 5(1), 1-6.

Brodie, T. A., Holmes, P. H., & Urquhart, G. M. (1986). Some aspects of tick-borne diseases of British sheep. *The Veterinary Record*, 118(15), 415-418.

Brown, A. C., Bryant, J. M., Einer-Jensen, K., Holdstock, J., Houniet, D. T., Chan, J. Z., ... & Breuer, J. (2015). Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *Journal of clinical microbiology*, 53(7), 2230-2237.

Cabezas-Cruz, A., Espinosa, P. J., Obregón, D. A., Alberdi, P., & de La Fuente, J. (2017). *Ixodes scapularis* tick cells control *Anaplasma phagocytophilum* infection by increasing the synthesis of phosphoenolpyruvate from tyrosine. *Frontiers in Cellular and Infection Microbiology*, 7, 375.

Cai, W., Nunziata, S., Rascoe, J., & Stulberg, M. J. (2019). SureSelect targeted enrichment, a new cost effective method for the whole genome sequencing of *Candidatus Liberibacter asiaticus*. *Scientific reports*, 9(1), 18962.

Centre for Tropical Veterinary Medicine. (1979). Annual Report. University of Edinburgh.

Centre for Tropical Veterinary Medicine. (1982). Annual Report. University of Edinburgh.

Centre for Tropical Veterinary Medicine. (1983). Annual Report. University of Edinburgh.

Chastagner, A., Dugat, T., Vourc'h, G., Verheyden, H., Legrand, L., Bachy, V., ... & Leblond, A. (2014). Multilocus sequence analysis of *Anaplasma phagocytophilum* reveals three distinct lineages with different host ranges in clinically ill French cattle. *Veterinary research*, 45(1), 1-12.

Chastagner, A., Pion, A., Verheyden, H., Lourtet, B., Cargnelutti, B., Picot, D., ... & Bailly, X. (2017). Host specificity, pathogen exposure, and superinfections impact the distribution of *Anaplasma phagocytophilum* genotypes in ticks, roe deer, and livestock in a fragmented agricultural landscape. *Infection, Genetics and Evolution*, 55, 31-44.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PloS one*, 8(4), e62856.

Choat, J., Yockey, B., Sheldon, S. W., Pappert, R., Petersen, J., & Dietrich, E. A. (2023). Development and Validation of a Real-Time PCR Test to Detect *Bartonella quintana* in Clinical Samples. *Diagnostic Microbiology and Infectious Disease*, 116000.

Chomel, B. B., & Kasten, R. W. (2010). Bartonellosis, an increasingly recognized zoonosis. *Journal of Applied Microbiology*, 109(3), 743-750.

Chomel, B. B., Boulouis, H. J., Breitschwerdt, E. B., Kasten, R. W., Vayssier-Taussat, M., Birtles, R. J., ... & Dehio, C. (2009). Ecological fitness and strategies of adaptation of *Bartonella* species to their hosts and vectors. *Veterinary research*, 40(2).

Christiansen, M. T., Brown, A. C., Kundu, S., Tutill, H. J., Williams, R., Brown, J. R., ... & Breuer, J. (2014). Whole-genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples. *BMC infectious diseases*, 14, 1-11.

Christie, P. J. (2004). Type IV secretion: the Agrobacterium VirB/D4 and related conjugation systems. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1694(1-3), 219-234.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., ... & Trujillo, M. E. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International journal of systematic and evolutionary microbiology*, 68(1), 461-466.

Clark, S. A., Doyle, R., Lucidarme, J., Borrow, R., & Breuer, J. (2018). Targeted DNA enrichment and whole genome sequencing of Neisseria meningitidis directly from clinical specimens. *International Journal of Medical Microbiology*, 308(2), 256-262.

Colborn, J. M., Kosoy, M. Y., Motin, V. L., Telepnev, M. V., Valbuena, G., Myint, K. S., ... & Peruski, L. (2010). Improved detection of *Bartonella* DNA in mammalian hosts and arthropod vectors by real-time PCR using the NADH dehydrogenase gamma subunit (nuoG). *Journal of clinical microbiology*, 48(12), 4630-4633.

Colwell, R. R. (1970). Polyphasic taxonomy of the genus Vibrio: numerical taxonomy of Vibrio cholerae, Vibrio parahaemolyticus, and related Vibrio species. *Journal of Bacteriology*, 104(1), 410-433.

- Courtney, J. W., Kostelnik, L. M., Zeidner, N. S., & Massung, R. F. (2004). Multiplex real-time PCR for detection of *Anaplasma phagocytophilum* and *Borrelia burgdorferi*. *Journal of Clinical Microbiology*, 42(7), 3164-3168.
- Crosby, F. L., Eskeland, S., Bø-Granquist, E. G., Munderloh, U. G., Price, L. D., Al-Khedery, B., ... & Barbet, A. F. (2022). Comparative whole genome analysis of an *Anaplasma phagocytophilum* strain isolated from Norwegian sheep. *Pathogens*, 11(5), 601.
- Crosby, F. L., Munderloh, U. G., Nelson, C. M., Herron, M. J., Lundgren, A. M., Xiao, Y. P., ... & Barbet, A. F. (2020). Disruption of VirB6 paralogs in *Anaplasma phagocytophilum* attenuates its growth. *Journal of bacteriology*, 202(23), 10-1128.
- Cross, A. S. (2008). What is a virulence factor?. *Critical Care*, 12(6), 1-2.
- Dahlgren, F. S., Mandel, E. J., Krebs, J. W., Massung, R. F., & McQuiston, J. H. (2011). Increasing incidence of *Ehrlichia chaffeensis* and *Anaplasma phagocytophilum* in the United States, 2000–2007. *The American journal of tropical medicine and hygiene*, 85(1), 124.
- Dantas-Torres, F., & Otranto, D. (2017). Anaplasmosis. *Arthropod borne diseases*, 215-222.
- Darby, A. C., Cho, N. H., Fuxelius, H. H., Westberg, J., & Andersson, S. G. (2007). Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *TRENDS in Genetics*, 23(10), 511-520.
- de Albuquerque, G. E., Moda, B. S., Serpa, M. S., Branco, G. P., Defelicibus, A., Takenaka, I. K., ... & Dias-Neto, E. (2022). Evaluation of bacteria and fungi DNA abundance in human tissues. *Genes*, 13(2), 237.
- De La Fuente, J., Massung, R. F., Wong, S. J., Chu, F. K., Lutz, H., Meli, M., ... & Kocan, K. M. (2005). Sequence analysis of the *msp4* gene of *Anaplasma phagocytophilum* strains. *Journal of Clinical Microbiology*, 43(3), 1309-1317.
- de Sá, P. H., Guimarães, L. C., das Graças, D. A., de Oliveira Veras, A. A., Barh, D., Azevedo, V., ... & Ramos, R. T. (2018). Next-generation sequencing and data analysis: strategies, tools, pipelines and protocols. In *Omics Technologies and Bio-Engineering* (pp. 191-207). Academic Press.
- Deng, H. K., Cescau, S., Danchin, A., Yang, H., Quebatte, M., Engel, P., ... & Dehio, C. (2010). The Trw Type IV Secretion System of *Bartonella* Mediates Host-Specific Adhesion to Erythrocytes.
- Deng, H. K., Le Rhun, D., Le Naour, E., Bonnet, S., & Vayssier-Taussat, M. (2012). Identification of *Bartonella* Trw host-specific receptor on erythrocytes.
- Deng, H., Le Rhun, D., Buffet, J. P. R., Cotté, V., Read, A., Birtles, R. J., & Vayssier-Taussat, M. (2012). Strategies of exploitation of mammalian reservoirs by *Bartonella* species. *Veterinary research*, 43, 1-14.
- Deng, H., Wu, S., Song, Q., Zhang, J., Sang, F., Sun, X., ... & Zhao, B. (2019). Cloning and identification of *Bartonella*  $\alpha$ -enolase as a plasminogen-binding protein. *Microbial pathogenesis*, 135, 103651.
- Dennis, T. P., Mable, B. K., Brunelle, B., Devault, A., Carter, R. W., Ling, C. L., ... & Forde, T. L. (2022). Target-enrichment sequencing yields valuable genomic data for challenging-to-culture bacteria of public health importance. *Microbial Genomics*, 8(5), 000836.



Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y. C., Gray, E. R., Grant, P., ... & Breuer, J. (2011). Specific capture and whole-genome sequencing of viruses from clinical samples. *PloS one*, 6(11), e27805.

Diaz, M. H., Bai, Y., Malania, L., Winchell, J. M., & Kosoy, M. Y. (2012). Development of a novel genus-specific real-time PCR assay for detection and differentiation of *Bartonella* species and genotypes. *Journal of clinical microbiology*, 50(5), 1645-1649.

Dicenzo, G. C., Mengoni, A., & Perrin, E. (2019). Chromids aid genome expansion and functional diversification in the family Burkholderiaceae. *Molecular biology and evolution*, 36(3), 562-574.

Diop, A., Raoult, D., & Fournier, P. E. (2018). Rickettsial genomics and the paradigm of genome reduction associated with increased virulence. *Microbes and Infection*, 20(7-8), 401-409.

Diuk-Wasser, M. A., & EPJ, V. (2016). Krause Coinfection by the tick-borne pathogens *Babesia microti* and *Borrelia burgdorferi*: ecological, epidemiological and clinical consequences. *Trends Parasitol*, 32(1), 30-42.

Doyle, R. M., Burgess, C., Williams, R., Gorton, R., Booth, H., Brown, J., ... & Breuer, J. (2018). Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *Journal of Clinical Microbiology*, 56(8), 10-1128.

Drazenovich, N., Foley, J., & Brown, R. N. (2006). Use of real-time quantitative PCR targeting the *msp2* protein gene to identify cryptic *Anaplasma phagocytophilum* infections in wildlife and domestic animals. *Vector-Borne & Zoonotic Diseases*, 6(1), 83-90.

Duan, R., Lv, D., Fan, R., Fu, G., Mu, H., Xi, J., ... & Wang, X. (2022). *Anaplasma phagocytophilum* in *Marmota himalayana*. *BMC genomics*, 23(1), 335.

Dugat, T., Chastagner, A., Lagrée, A. C., Petit, E., Durand, B., Thierry, S., ... & Haddad, N. (2014). A new multiple-locus variable-number tandem repeat analysis reveals different clusters for *Anaplasma phagocytophilum* circulating in domestic and wild ruminants. *Parasites & vectors*, 7(1), 1-11.

Dugat, T., Lagrée, A. C., Maillard, R., Boulouis, H. J., & Haddad, N. (2015). Opening the black box of *Anaplasma phagocytophilum* diversity: current situation and future perspectives. *Frontiers in cellular and infection microbiology*, 5, 61.

Dumler, J. S., Barbet, A. F., Bekker, C. P., Dasch, G. A., Palmer, G. H., Ray, S. C., ... & Rurangirwa, F. R. (2001). Reorganization of genera in the families Rickettsiaceae and Anaplasmataceae in the order Rickettsiales: unification of some species of *Ehrlichia* with *Anaplasma*, *Cowdria* with *Ehrlichia* and *Ehrlichia* with *Neorickettsia*, descriptions of six new species combinations and designation of *Ehrlichia equi* and 'HGE agent' as subjective synonyms of *Ehrlichia phagocytophila*. *International journal of systematic and evolutionary microbiology*, 51(6), 2145-2165.

Dumler, J. S., Choi, K. S., Garcia-Garcia, J. C., Barat, N. S., Scorpio, D. G., Garyu, J. W., ... & Bakken, J. S. (2005). Human granulocytic anaplasmosis and *Anaplasma phagocytophilum*. *Emerging infectious diseases*, 11(12), 1828.

Engel, P., Salzburger, W., Liesch, M., Chang, C. C., Maruyama, S., Lanz, C., ... & Dehio, C. (2011). Parallel evolution of a type IV secretion system in radiating lineages of the host-restricted bacterial pathogen *Bartonella*. *PLoS Genet*, 7(2), e1001296.

- Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3, e1319.
- Ernst, E., Qurollo, B., Olech, C., & Breitschwerdt, E. B. (2020). *Bartonella rochalimae*, a newly recognized pathogen in dogs. *Journal of Veterinary Internal Medicine*, 34(4), 1447-1453.
- Feehery, G. R., Yigit, E., Oyola, S. O., Langhorst, B. W., Schmidt, V. T., Stewart, F. J., ... & Pradhan, S. (2013). A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PloS one*, 8(10), e76096.
- Ferrolho, J., Simpson, J., Hawes, P., Zweggarth, E., & Bell-Sakyi, L. (2016). Growth of *Ehrlichia canis*, the causative agent of canine monocytic ehrlichiosis, in vector and non-vector ixodid tick cell lines. *Ticks and tick-borne diseases*, 7(4), 631-637.
- Finkelstein, J. L., Brown, T. P., O'reilly, K. L., Wedincamp Jr, J., & Foil, L. D. (2002). Studies on the growth of *Bartonella henselae* in the cat flea (Siphonaptera: Pulicidae). *Journal of medical entomology*, 39(6), 915-919.
- Foley, J. E., Nieto, N. C., Adjemian, J., Dabritz, H., & Brown, R. N. (2008). *Anaplasma phagocytophilum* infection in small mammal hosts of *Ixodes* ticks, western United States. *Emerging Infectious Diseases*, 14(7), 1147.
- Fondi, M., Emiliani, G., & Fani, R. (2009). Origin and evolution of operons and metabolic pathways. *Research in Microbiology*, 160(7), 502-512.
- Foster, W. N. M., & Cameron, A. E. (1970). I Observations on ovine strains of tick-borne fever. *Journal of comparative pathology*, 80, 429-436.
- Fromm, K., & Dehio, C. (2021). The impact of *Bartonella* VirB/VirD4 type IV secretion system effectors on eukaryotic host cells. *Frontiers in Microbiology*, 12, 762582.
- Fromm, K., Ortell, M., Boegli, A., & Dehio, C. (2024). Translocation of YopJ family effector proteins through the VirB/VirD4 T4SS of *Bartonella*. *Proceedings of the National Academy of Sciences*, 121(20), e2310348121.
- Gai, M., d'Onofrio, G., di Vico, M. C., Ranghino, A., Nappo, A., Diena, D., ... & Biancone, L. (2015, September). Cat-scratch disease: case report and review of the literature. In *Transplantation Proceedings* (Vol. 47, No. 7, pp. 2245-2247). Elsevier.
- Gandy, S., Hansford, K., McGinley, L., Cull, B., Smith, R., Semper, A., ... & Medlock, J. M. (2022). Prevalence of *Anaplasma phagocytophilum* in questing *Ixodes ricinus* nymphs across twenty recreational areas in England and Wales. *Ticks and Tick-borne Diseases*, 13(4), 101965.
- Gaudin, M., & Desnues, C. (2018). Hybrid capture-based next generation sequencing and its application to human infectious diseases. *Frontiers in microbiology*, 9, 2924.
- Ghafar, M. W., & Amer, S. A. (2012). Prevalence and first molecular characterization of *Anaplasma phagocytophilum*, the agent of human granulocytic anaplasmosis, in *Rhipicephalus sanguineus* ticks attached to dogs from Egypt. *Journal of Advanced Research*, 3(2), 189-194.
- Gilbert, L., Maffey, G. L., Ramsay, S. L., & Hester, A. J. (2012). The effect of deer management on the abundance of *Ixodes ricinus* in Scotland. *Ecological Applications*, 22(2), 658-667.

- Gilchrist, C. L., & Chooi, Y. H. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, 37(16), 2473-2475.
- Gillespie, J. J., Brayton, K. A., Williams, K. P., Quevedo Diaz, M. A., Brown, W. C., Azad, A. F., & Sobral, B. W. (2010). Phylogenomics reveals a diverse Rickettsiales type IV secretion system. *Infection and immunity*, 78(5), 1809-1823.
- Gonçalves-Oliveira, J., Gutierrez, R., Schlesener, C. L., Jaffe, D. A., Aguilar-Setién, A., Boulouis, H. J., ... & Harrus, S. (2023). Genomic characterization of three novel *Bartonella* strains in a rodent and two bat species from Mexico. *Microorganisms*, 11(2), 340.
- Goodman, J. L., Nelson, C., Vitale, B., Madigan, J. E., Dumler, J. S., Kurtti, T. J., & Munderloh, U. G. (1996). Direct cultivation of the causative agent of human granulocytic ehrlichiosis. *New England Journal of Medicine*, 334(4), 209-215.
- Gordon, W. S., Brownlee, A., Wilson, D. R., & MacLeod, J. (1932). "Tick-borne Fever"(A hitherto undescribed Disease of Sheep). *Journal of Comparative Pathology*, 45(pt. 4).
- Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, 57(1), 81-91.
- Grassi, L., Franzo, G., Martini, M., Mondin, A., Cassini, R., Drigo, M., ... & Menandro, M. L. (2021). Ecotyping of *Anaplasma phagocytophilum* from wild ungulates and ticks shows circulation of zoonotic strains in northeastern Italy. *Animals*, 11(2), 310.
- Guptill, L. (2010). Bartonellosis. *Veterinary microbiology*, 140(3-4), 347-359.
- Guterres, A., Gonçalves, J., & de Lemos, E. R. S. (2019). What is the minimum length of gltA gene required for phylogenetic analyzes in *Bartonella*? *Research in microbiology*, 170(1), 60-64.
- Gutiérrez, R., Shalit, T., Markus, B., Yuan, C., Nachum-Biala, Y., Elad, D., & Harrus, S. (2020). *Bartonella kosoyi* sp. nov. and *Bartonella krasnovii* sp. nov., two novel species closely related to the zoonotic *Bartonella elizabethae*, isolated from black rats and wild desert rodent-fleas. *International journal of systematic and evolutionary microbiology*, 70(3), 1656-1665.
- Guy, L., Nystedt, B., Toft, C., Zaremba-Niedzwiedzka, K., Berglund, E. C., Granberg, F., ... & Andersson, S. G. (2013). A gene transfer agent and a dynamic repertoire of secretion systems hold the keys to the explosive radiation of the emerging pathogen *Bartonella*. *PLoS Genetics*, 9(3), e1003393.
- Han, H. J., Wen, H. L., Zhao, L., Liu, J. W., Luo, L. M., Zhou, C. M., ... & Yu, X. J. (2017). Novel *Bartonella* species in insectivorous bats, northern China. *PLoS One*, 12(1), e0167915.
- Harms, A., & Dehio, C. (2012). Intruders below the radar: molecular pathogenesis of *Bartonella* spp. *Clinical microbiology reviews*, 25(1), 42-78.
- Harms, A., Liesch, M., Körner, J., Québatte, M., Engel, P., & Dehio, C. (2017). A bacterial toxin-antitoxin module is the origin of inter-bacterial and inter-kingdom effectors of *Bartonella*. *PLoS genetics*, 13(10), e1007077.

Harms, A., Segers, F. H., Quebatte, M., Mistl, C., Manfredi, P., Körner, J., ... & Dehio, C. (2017). Evolutionary dynamics of pathoadaptation revealed by three independent acquisitions of the VirB/D4 type IV secretion system in *Bartonella*. *Genome biology and evolution*, 9(3), 761-776.

Harms, A., Stanger, F. V., & Dehio, C. (2016). Biological diversity and molecular plasticity of FIC domain proteins. *Annual review of microbiology*, 70, 341-360.

Harms, A., Stanger, F. V., Scheu, P. D., de Jong, I. G., Goepfert, A., Glatter, T., ... & Dehio, C. (2015). Adenylation of gyrase and topo IV by FicT toxins disrupts bacterial DNA topology. *Cell reports*, 12(9), 1497-1507.

Hawlana, H., Rynkiewicz, E., Toh, E., Alfred, A., Durden, L. A., Hastriter, M. W., ... & Clay, K. (2013). The arthropod, but not the vertebrate host or its environment, dictates bacterial community composition of fleas and ticks. *The ISME journal*, 7(1), 221-223.

Hayashi Sant'Anna, F., Bach, E., Porto, R. Z., Guella, F., Hayashi Sant'Anna, E., & Passaglia, L. M. (2019). Genomic metrics made easy: what to do and where to go in the new era of bacterial taxonomy. *Critical reviews in microbiology*, 45(2), 182-200.

Headquarters, M. A., West, A., Centers–Boston, A., & Farm, A. C. N. (2016). Feline Anaplasmosis. *Internal Medicine*, 1, 131.

Hendrix, L. R. (2000). Contact-dependent hemolytic activity distinct from deforming activity of *Bartonella bacilliformis*. *FEMS microbiology letters*, 182(1), 119-124.

Hendrix, L. R., & Kiss, K. (2003). Studies on the identification of deforming factor from *Bartonella bacilliformis*. *Annals of the New York Academy of Sciences*, 990(1), 596-604.

Henn, J. B., Chomel, B. B., Boulouis, H. J., Kasten, R. W., Murray, W. J., Bar-Gal, G. K., ... & Baneth, G. (2009). *Bartonella rochalimae* in raccoons, coyotes, and red foxes. *Emerging infectious diseases*, 15(12), 1984.

Hoarau, A. O., Mavingui, P., & Lebarbenchon, C. (2020). Coinfections in wildlife: Focus on a neglected aspect of infectious disease epidemiology. *PLoS Pathogens*, 16(9), e1008790.

Hofmeester, T. R., Jansen, P. A., Wijnen, H. J., Coipan, E. C., Fonville, M., Prins, H. H., ... & van Wieren, S. E. (2017). Cascading effects of predator activity on tick-borne disease risk. *Proceedings of the Royal Society B: Biological Sciences*, 284(1859), 20170453.

Holman, P. J. (1981). Partial characterization of a unique female diploid cell strain from the tick *Boophilus microplus* (Acari: Ixodidae). *Journal of Medical Entomology*, 18(1), 84-88.

Holman, P. J., & Onald, N. C. R. (1980). A new tick cell line derived from *Boophilus microplus*. *Research in veterinary science*, 29(3), 383-387.

Holmberg, M., Mills, J. N., McGill, S., Benjamin, G., & Ellis, B. A. (2003). *Bartonella* infection in sylvatic small mammals of central Sweden. *Epidemiology & Infection*, 130(1), 149-157.

Hoppner, C., Carle, A., Sivanesan, D., Hoepfner, S., & Baron, C. (2005). The putative lytic transglycosylase VirB1 from *Brucella suis* interacts with the type IV secretion system core components VirB8, VirB9 and VirB11. *Microbiology*, 151(11), 3469-3482.

- Horsfield, S. T., Tonkin-Hill, G., Croucher, N. J., & Lees, J. A. (2023). Accurate and fast graph-based pangenome annotation and clustering with ggCaller. *Genome Research*, 33(9), 1622-1637.
- Huang, K. Y. (1967). Metabolic activity of the trench fever rickettsia, *Rickettsia quintana*. *Journal of Bacteriology*, 93(3), 853-859.
- Huhn, C., Winter, C., Wolfsperger, T., Wüppenhorst, N., Strašek Smrdel, K., Skuballa, J., ... & Von Loewenich, F. D. (2014). Analysis of the population structure of *Anaplasma phagocytophilum* using multilocus sequence typing. *PLoS One*, 9(4), e93725.
- Hulínská, D., LANGROVÁ, K., PEJČOCH, M., & Pavlásek, I. (2004). Detection of *Anaplasma phagocytophilum* in animals by real-time polymerase chain reaction. *Apmis*, 112(4-5), 239-247.
- Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, 1-11.
- IJdo, J. W., Zhang, Y., Hodzic, E., Magnarelli, L. A., Wilson, M. L., Telford III, S. R., ... & Fikrig, E. (1997). The early humoral response in human granulocytic ehrlichiosis. *Journal of Infectious Diseases*, 176(3), 687-692.
- Inoue, K., Maruyama, S., Kabeya, H., Hagiya, K., Izumi, Y., Une, Y., & Yoshikawa, Y. (2009). Exotic small mammals as potential reservoirs of zoonotic *Bartonella* spp. *Emerging Infectious Diseases*, 15(4), 526.
- Jaarsma, R. I., Sprong, H., Takumi, K., Kazimirova, M., Silaghi, C., Myrsterud, A., ... & Estrada-Peña, A. (2019). *Anaplasma phagocytophilum* evolves in geographical and biotic niches of vertebrates and ticks. *Parasites & vectors*, 12, 1-17.
- Jahfari, S., Coipan, E. C., Fonville, M., Van Leeuwen, A. D., Hengeveld, P., Heylen, D., ... & Sprong, H. (2014). Circulation of four *Anaplasma phagocytophilum* ecotypes in Europe. *Parasites & vectors*, 7, 1-11.
- Janda, J. M. (2018). Clinical decisions: how relevant is modern bacterial taxonomy for clinical microbiologists?. *Clinical Microbiology Newsletter*, 40(7), 51-57.
- Jin, H., Wei, F., Liu, Q., & Qian, J. (2012). Epidemiology and control of human granulocytic anaplasmosis: a systematic review. *Vector-Borne and Zoonotic Diseases*, 12(4), 269-274.
- Jin, X., Gou, Y., Xin, Y., Li, J., Sun, J., Li, T., & Feng, J. (2023). Advancements in understanding the molecular and immune mechanisms of *Bartonella* pathogenicity. *Frontiers in Microbiology*, 14, 1196700.
- Johnson, N., Golding, M., & Phipps, L. P. (2021). Detection of Tick-Borne Pathogens in Red Deer (*Cervus elaphus*), United Kingdom. *Pathogens*, 10(6), 640.
- Jones, R. T., McCormick, K. F., & Martin, A. P. (2008). Bacterial communities of *Bartonella*-positive fleas: diversity and community assembly patterns. *Applied and environmental microbiology*, 74(5), 1667-1670.
- Joseph, S. J., Cox, D., Wolff, B., Morrison, S. S., Kozak-Muiznieks, N. A., Frace, M., ... & Dean, D. (2016). Dynamics of genome change among *Legionella* species. *Scientific reports*, 6(1), 33442.
- Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: how whole-genome sequencing and big data are transforming epidemiology. *Trends in microbiology*, 22(5), 282-291.

Karshima, S. N., Ahmed, M. I., Kogi, C. A., & Iliya, P. S. (2022). *Anaplasma phagocytophilum* infection rates in questing and host-attached ticks: a global systematic review and meta-analysis. *Acta Tropica*, 228, 106299.

Keebaugh, E. S., & Schlenke, T. A. (2014). Insights from natural host–parasite interactions: The *Drosophila* model. *Developmental & Comparative Immunology*, 42(1), 111-123.

Kelly, T. M., Padmalayam, I., & Baumstark, B. R. (1998). Use of the cell division protein FtsZ as a means of differentiating among *Bartonella* species. *Clinical Diagnostic Laboratory Immunology*, 5(6), 766-772.

Keragala, C. B., & Medcalf, R. L. (2021). Plasminogen: an enigmatic zymogen. *Blood, The Journal of the American Society of Hematology*, 137(21), 2881-2889.

Kim, M., Oh, H. S., Park, S. C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International journal of systematic and evolutionary microbiology*, 64(Pt\_2), 346-351.

Knobloch, J., Bialek, R., Müller, G., & Asmus, P. (1988). Common surface epitope of *Bartonella bacilliformis* and *Chlamydia psittaci*. *The American journal of tropical medicine and hygiene*, 39(5), 427-433.

Koesling, J., Aebischer, T., Falch, C., Schüle, R., & Dehio, C. (2001). Cutting edge: antibody-mediated cessation of hemotropic infection by the intraerythrocytic mouse pathogen *Bartonella grahamii*. *The Journal of Immunology*, 167(1), 11-14.

Kosoy, M. Y., Regnery, R. L., Kosaya, O. I., Jones, D. C., Marston, E. L., & Childs, J. E. (1998). Isolation of *Bartonella* spp. from embryos and neonates of naturally infected rodents. *Journal of Wildlife Diseases*, 34(2), 305-309.

Kosoy, M., Hayman, D. T., & Chan, K. S. (2012). *Bartonella* bacteria in nature: where does population variability end and a species start?. *Infection, Genetics and Evolution*, 12(5), 894-904.

Krawczyk, A. I., Röttgers, S., Coimbra-Dores, M. J., Heylen, D., Fonville, M., Takken, W., ... & Sprong, H. (2022). Tick microbial associations at the crossroad of horizontal and vertical transmission pathways. *Parasites & vectors*, 15(1), 380.

Kreier, J. P., Gothe, R., Ihler, G. M., Krampitz, H. E., Mernaugh, G., & Palmer, G. H. (1992). The hemotropic bacteria: the families Bartonellaceae and Anaplasmataceae. In *The Prokaryotes: a Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications* (pp. 3994-4022). New York, NY: Springer New York.

Krügel, M., Król, N., Kempf, V. A., Pfeffer, M., & Obiegala, A. (2022). Emerging rodent-associated *Bartonella*: a threat for human health?. *Parasites & Vectors*, 15(1), 113.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9), 1639-1645.

Kumari, R., Shariq, M., Sharma, S., Kumar, A., & Mukhopadhyay, G. (2019). CagW, a VirB6 homologue interacts with Cag-type IV secretion system substrate CagA in *Helicobacter pylori*. *Biochemical and biophysical research communications*, 515(4), 712-718.

- Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2021). An update for taxonomy designers: methodological guidance from information systems research. *Business & Information Systems Engineering*, 1-19.
- Kuo, C. C., Huang, J. L., Chien, C. H., Shih, H. C., & Wang, H. C. (2018). First molecular detection of *Anaplasma phagocytophilum* in the hard tick *Rhipicephalus haemaphysaloides* in Taiwan. *Experimental and Applied Acarology*, 75, 437-443.
- Kurtti, T. J., & Munderloh, U. G. (1982). Tick cell culture: characteristics, growth requirements, and applications to parasitology. In *Invertebrate cell culture applications* (pp. 195-232). Academic press.
- Kurtti, T. J., Munderloh, U. G., & Samish, M. (1982). Effect of medium supplements on tick cells in culture. *The Journal of parasitology*, 930-935.
- Kurtti, T. J., Munderloh, U. G., & Stiller, D. (1983). The interaction of *Babesia caballi* kinetes with tick cells. *Journal of invertebrate pathology*, 42(3), 334-343.
- Kurtti, T. J., Munderloh, U. G., Andreadis, T. G., Magnarelli, L. A., & Mather, T. N. (1996). Tick cell culture isolation of an intracellular Prokaryote from the Tick *Ixodes scapularis*. *Journal of invertebrate pathology*, 67(3), 318-321.
- Kyme, P., Dillon, B., & Iredell, J. (2003). Phase variation in *Bartonella henselae*. *Microbiology*, 149(3), 621-629.
- La Scola, B., Zeaiter, Z., Khamis, A., & Raoult, D. (2003). Gene-sequence-based criteria for species definition in bacteriology: the *Bartonella* paradigm. *Trends in microbiology*, 11(7), 318-321.
- Lähteenmäki, K., Kuusela, P., & Korhonen, T. K. (2001). Bacterial plasminogen activators and receptors. *FEMS microbiology reviews*, 25(5), 531-552.
- Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The distribution, evolution, and roles of gene transfer agents in prokaryotic genetic exchange. *Annual review of virology*, 4, 87-104.
- Langenwalder, D. B., Schmidt, S., Silaghi, C., Skuballa, J., Pantchev, N., Matei, I. A., ... & von Loewenich, F. D. (2020). The absence of the *drhm* gene is not a marker for human-pathogenicity in European *Anaplasma phagocytophilum* strains. *Parasites & vectors*, 13(1), 1-13.
- Larrea, D., De Paz, H. D., Arechaga, I., de la Cruz, F., & Llosa, M. (2013). Structural independence of conjugative coupling protein TrwB from its type IV secretion machinery. *Plasmid*, 70(1), 146-153.
- Lee, S. H., Shin, N. R., Kim, C. M., Park, S., Yun, N. R., Kim, D. M., & Jung, D. S. (2020). First identification of *Anaplasma phagocytophilum* in both a biting tick *Ixodes nipponensis* and a patient in Korea: A case report. *BMC Infectious Diseases*, 20, 1-10.
- Li, D. M., Hou, Y., Song, X. P., Fu, Y. Q., Li, G. C., Li, M., ... & Liu, Q. Y. (2015). High prevalence and genetic heterogeneity of rodent-borne *Bartonella* species on Heixiazhi Island, China. *Applied and Environmental Microbiology*, 81(23), 7981-7992.
- Li, R., Ma, Z., Zheng, W., Wang, Z., Yi, J., Xiao, Y., ... & Chen, C. (2022). Multiomics analyses reveals *Anaplasma phagocytophilum* Ats-1 induces anti-apoptosis and energy metabolism by upregulating the respiratory chain-mPTP axis in eukaryotic mitochondria. *BMC microbiology*, 22(1), 271.

- Liberto, M. C., & Matera, G. (2000). Pathogenic mechanisms of *Bartonella quintana*. *The New Microbiologica*, 23(4), 449-456.
- Lin, J. W., Hsu, Y. M., Chomel, B. B., Lin, L. K., Pei, J. C., Wu, S. H., & Chang, C. C. (2012). Identification of novel *Bartonella* spp. in bats and evidence of Asian gray shrew as a new potential reservoir of *Bartonella*. *Veterinary microbiology*, 156(1-2), 119-126.
- Lin, M., & Rikihisa, Y. (2003). Ehrlichia chaffeensis and *Anaplasma phagocytophilum* lack genes for lipid A biosynthesis and incorporate cholesterol for their survival. *Infection and immunity*, 71(9), 5324-5331.
- Lin, Y., Zhang, Y., Sun, H., Jiang, H., Zhao, X., Teng, X., ... & Zhou, J. (2024). NanoDeep: a deep learning framework for nanopore adaptive sampling on microbial sequencing. *Briefings in Bioinformatics*, 25(1), bbad499.
- Litwin, C. M., & Johnson, J. M. (2005). Identification, cloning, and expression of the CAMP-like factor autotransporter gene (cfa) of *Bartonella henselae*. *Infection and immunity*, 73(7), 4205-4213.
- Llanes, A., & Rajeev, S. (2020). First whole genome sequence of *Anaplasma platys*, an obligate intracellular rickettsial pathogen of dogs. *Pathogens*, 9(4), 277.
- Maggi, R. G., Duncan, A. W., & Breitschwerdt, E. B. (2005). Novel chemically modified liquid medium that will support the growth of seven *Bartonella* species. *Journal of clinical microbiology*, 43(6), 2651-2655.
- Majazki, J., Wüppenhorst, N., Hartelt, K., Birtles, R., & Von Loewenich, F. D. (2013). *Anaplasma phagocytophilum* strains from voles and shrews exhibit specific ankA gene sequences. *BMC veterinary research*, 9, 1-7.
- Makarova, K. S., Aravind, L., Wolf, Y. I., Tatusov, R. L., Minton, K. W., Koonin, E. V., & Daly, M. J. (2001). Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiology and molecular biology reviews*, 65(1), 44-79.
- Malgorzata-Miller, G., Heinbockel, L., Brandenburg, K., van der Meer, J. W., Netea, M. G., & Joosten, L. A. (2016). *Bartonella quintana* lipopolysaccharide (LPS): structure and characteristics of a potent TLR4 antagonist for in-vitro and in-vivo applications. *Scientific reports*, 6(1), 34221.
- Malmsten, J., Widen, D. G., Rydevik, G., Yon, L., Hutchings, M. R., Thulin, C. G., ... & Dalin, A. M. (2014). Temporal and spatial variation in *Anaplasma phagocytophilum* infection in Swedish moose (*Alces alces*). *Epidemiology & Infection*, 142(6), 1205-1213.
- Mändle, T., Einsele, H., Schaller, M., Neumann, D., Vogel, W., Autenrieth, I. B., & Kempf, V. A. (2005). Infection of human CD34+ progenitor cells with *Bartonella henselae* results in intraerythrocytic presence of *B. henselae*. *Blood*, 106(4), 1215-1222.
- Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., & Leggett, R. M. (2022). Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome biology*, 23(1), 11.
- Mason, R. A. (1970). Propagation and growth cycle of *Rickettsia quintana* in a new liquid medium. *Journal of Bacteriology*, 103(1), 184-190.



- Massung, R. F., Levin, M. L., Munderloh, U. G., Silverman, D. J., Lynch, M. J., Gaywee, J. K., & Kurtti, T. J. (2007). Isolation and propagation of the Ap-Variant 1 strain of *Anaplasma phagocytophilum* in a tick cell line. *Journal of clinical microbiology*, 45(7), 2138-2143.
- Massung, R. F., Mauel, M. J., Owens, J. H., Allan, N., Courtney, J. W., Stafford III, K. C., & Mather, T. N. (2002). Genetic variants of Ehrlichia phagocytophila, Rhode Island and Connecticut. *Emerging Infectious Diseases*, 8(5), 467.
- Matei, I. A., Estrada-Peña, A., Cutler, S. J., Vayssier-Taussat, M., Varela-Castro, L., Potkonjak, A., ... & Mihalca, A. D. (2019). A review on the eco-epidemiology and clinical management of human granulocytic anaplasmosis and its agent in Europe. *Parasites & vectors*, 12, 1-19.
- Matera, G., Liberto, M. C., Joosten, L. A., Vinci, M., Quirino, A., Pulicari, M. C., ... & Focà, A. (2008). The Janus face of *Bartonella quintana* recognition by Toll-like receptors (TLRs): a review. *European cytokine network*, 19(3), 113-118.
- McCann, C. D., & Jordan, J. A. (2014). Evaluation of MolYsis™ Complete5 DNA extraction method for detecting Staphylococcus aureus DNA from whole blood in a sepsis model using PCR/pyrosequencing. *Journal of microbiological methods*, 99, 1-7.
- McCutcheon, J. P., & Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1), 13-26.
- McGill, S. L., Regnery, R. L., & Karem, K. L. (1998). Characterization of human immunoglobulin (Ig) isotype and IgG subclass response to *Bartonella henselae* infection. *Infection and immunity*, 66(12), 5915-5920.
- Meziti, A., Rodriguez-R, L. M., Hatt, J. K., Peña-Gonzalez, A., Levy, K., & Konstantinidis, K. T. (2021). The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Applied and environmental microbiology*, 87(6), e02593-20.
- Michel, A. O., Mathis, A., & Ryser-Degiorgis, M. P. (2014). Babesia spp. in European wild ruminant species: parasite diversity and risk factors for infection. *Veterinary research*, 45, 1-11.
- Minnick, M. F. (1994). Identification of outer membrane proteins of *Bartonella bacilliformis*. *Infection and immunity*, 62(6), 2644-2648.
- Minnick, M. F., & Anderson, B. E. (2015). *Bartonella*. In *Molecular medical microbiology* (pp. 1911-1939). Academic Press.
- Minnick, M. F., & Barbian, K. D. (1997). Identification of *Bartonella* using PCR; genus-and species-specific primer sets. *Journal of microbiological methods*, 31(1-2), 51-57.
- Mitchell, S. J., & Minnick, M. F. (1995). Characterization of a two-gene locus from *Bartonella bacilliformis* associated with the ability to invade human erythrocytes. *Infection and Immunity*, 63(4), 1552-1562.
- Morick, D., Krasnov, B. R., Khokhlova, I. S., Gottlieb, Y., & Harrus, S. (2013). Transmission dynamics of *Bartonella* sp. strain OE 1-1 in Sundevall's jirds (*Meriones crassus*). *Applied and Environmental Microbiology*, 79(4), 1258-1264.

- Morway, C., Kosoy, M., Eisen, R., Montenieri, J., Sheff, K., Reynolds, P. J., & Powers, N. (2008). A longitudinal study of *Bartonella* infection in populations of woodrats and their fleas. *Journal of Vector Ecology*, 33(2), 353-364.
- Mosepele, M., Mazo, D., & Cohn, J. (2012). *Bartonella* infection in immunocompromised hosts: immunology of vascular infection and vasoproliferation. *Journal of Immunology Research*, 2012.
- Mukhacheva, T. A., Shaikhova, D. R., & Kovalev, S. Y. (2019). Asian isolates of *Anaplasma phagocytophilum*: Multilocus sequence typing. *Ticks and tick-borne diseases*, 10(4), 775-780.
- Munderloh, U. G., & Kurtti, T. J. (1989). Formulation of medium for tick cell culture. *Experimental & applied acarology*, 7, 219-229.
- Munderloh, U. G., Jauron, S. D., Fingerle, V., Leitritz, L., Hayes, S. F., Hautman, J. M., ... & Goodman, J. L. (1999). Invasion and intracellular development of the human granulocytic ehrlichiosis agent in tick cell culture. *Journal of Clinical Microbiology*, 37(8), 2518-2524.
- Munderloh, U. G., Liu, Y., Wang, M., Chen, C., & Kurtti, T. J. (1994). Establishment, maintenance and description of cell lines from the tick *Ixodes scapularis*. *The Journal of parasitology*, 533-543.
- Musso, T., Badolato, R., Ravarino, D., Stornello, S., Panzanelli, P., Merlino, C., ... & Zucca, M. (2001). Interaction of *Bartonella henselae* with the murine macrophage cell line J774: infection and proinflammatory response. *Infection and Immunity*, 69(10), 5974-5980.
- Netea, M. G., Suttmoller, R., Hermann, C., Van der Graaf, C. A., Van der Meer, J. W., Van Krieken, J. H., ... & Kullberg, B. J. (2004). Toll-like receptor 2 suppresses immunity against *Candida albicans* through induction of IL-10 and regulatory T cells. *The Journal of Immunology*, 172(6), 3712-3718.
- Niu, H., Kozjak-Pavlovic, V., Rudel, T., & Rikihisa, Y. (2010). *Anaplasma phagocytophilum* Ats-1 is imported into host cell mitochondria and interferes with apoptosis induction. *PLoS pathogens*, 6(2), e1000774.
- Norman, A. F., Regnery, R., Jameson, P., Greene, C., & Krause, D. (1995). Differentiation of *Bartonella*-like isolates at the species level by PCR-restriction fragment length polymorphism in the citrate synthase gene. *Journal of clinical microbiology*, 33(7), 1797-1803.
- Obino, D., & Duménil, G. (2019). The many faces of bacterium-endothelium interactions during systemic infections. *Microbiology Spectrum*, 7(2), 10-1128.
- Olival, K. J., Dittmar, K., Bai, Y., Rostal, M. K., Lei, B. R., Daszak, P., & Kosoy, M. (2015). *Bartonella* spp. in a Puerto Rican bat community. *Journal of Wildlife Diseases*, 51(1), 274-278.
- Oliver, J. D., Chávez, A. S. O., Felsheim, R. F., Kurtti, T. J., & Munderloh, U. G. (2015). An *Ixodes scapularis* cell line with a predominantly neuron-like phenotype. *Experimental and Applied Acarology*, 66, 427-442.
- Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., ... & Stevens, R. L. (2023). Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic acids research*, 51(D1), D678-D689.
- Oney, K., Koo, M., Roy, C., Ren, S., Quorllo, B., Juhasz, N. B., ... & Diniz, P. P. (2021). Evaluation of a commercial microbial enrichment kit used prior DNA extraction to improve the molecular detection of vector-borne pathogens from naturally infected dogs. *Journal of Microbiological Methods*, 188, 106163.

Oren, A., Arahal, D. R., Göker, M., Moore, E. R., Rossello-Mora, R., & Sutcliffe, I. C. (Eds.). (2023). International code of nomenclature of prokaryotes. Prokaryotic code (2022 revision). *International Journal of Systematic and Evolutionary Microbiology*, 73(5a), 005585.

Ostfeld, R. S., Schaubert, E. M., Canham, C. D., Keesing, F., Jones, C. G., & Wolff, J. O. (2001). Effects of acorn production and mouse abundance on abundance and *Borrelia burgdorferi* infection prevalence of nymphal *Ixodes scapularis* ticks. *Vector borne and zoonotic diseases*, 1(1), 55-63.

Palomar, A. M., Portillo, A., Santibáñez, P., Mazuelas, D., Roncero, L., García-Álvarez, L., ... & Oteo, J. A. (2015). Detection of tick-borne *Anaplasma bovis*, *Anaplasma phagocytophilum* and *Anaplasma centrale* in Spain. *Medical and veterinary entomology*, 29(3), 349-353.

Paziewska, A., Harris, P. D., Zwolińska, L., Bajer, A., & Siński, E. (2011). Recombination within and between species of the alpha proteobacterium *Bartonella* infecting rodents. *Microbial Ecology*, 61, 134-145.

Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular ecology*, 19(24), 5332-5344.

Pilloux, L., Baumgartner, A., Jaton, K., Lienhard, R., Ackermann-Gäumann, R., Beuret, C., & Greub, G. (2019). Prevalence of *Anaplasma phagocytophilum* and *Coxiella burnetii* in *Ixodes ricinus* ticks in Switzerland: an underestimated epidemiologic risk. *New microbes and new infections*, 27, 22-26.

Pons, M. J., Gomes, C., Aguilar, R., Barrios, D., Aguilar-Luis, M. A., Ruiz, J., ... & Moncunill, G. (2017). Immunosuppressive and angiogenic cytokine profile associated with *Bartonella bacilliformis* infection in post-outbreak and endemic areas of Carrion's disease in Peru. *PLoS Neglected Tropical Diseases*, 11(6), e0005684.

Popa, C., Abdollahi-Roodsaz, S., Joosten, L. A., Takahashi, N., Sprong, T., Matera, G., ... & Netea, M. G. (2007). *Bartonella quintana* lipopolysaccharide is a natural antagonist of Toll-like receptor 4. *Infection and immunity*, 75(10), 4831-4837.

Pretorius, A. M., Beati, L., & Birtles, R. J. (2004). Diversity of bartonellae associated with small mammals inhabiting Free State province, South Africa. *International journal of systematic and evolutionary microbiology*, 54(6), 1959-1967.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*, 8(1), 12-24.

Pulliainen, A. T., Pielles, K., Brand, C. S., Hauert, B., Böhm, A., Quebatte, M., ... & Dehio, C. (2012). Bacterial effector binds host cell adenylyl cyclase to potentiate Gas-dependent cAMP production. *Proceedings of the National Academy of Sciences*, 109(24), 9581-9586.

Rar, V. A., Epikhina, T. I., Yakimenko, V. V., Malkova, M. G., Tancev, A. K., Bondarenko, E. I., ... & Tikunova, N. V. (2014). Genetic variability of *Anaplasma phagocytophilum* in ticks and voles from *Ixodes persulcatus*/*Ixodes trianguliceps* sympatric areas from Western Siberia, Russia. *Ticks and Tick-borne Diseases*, 5(6), 854-863.

Rar, V., & Golovljova, I. (2011). *Anaplasma*, Ehrlichia, and "Candidatus Neoehrlichia" bacteria: pathogenicity, biodiversity, and molecular genetic characteristics, a review. *Infection, Genetics and Evolution*, 11(8), 1842-1861.

Rar, V., Tkachev, S., & Tikunova, N. (2021). Genetic diversity of *Anaplasma* bacteria: Twenty years later. *Infection, Genetics and Evolution*, 91, 104833.

- Regier, Y., O'Rourke, F., & Kempf, V. A. (2016). *Bartonella* spp.-a chance to establish One Health concepts in veterinary and human medicine. *Parasites & vectors*, 9(1), 1-12.
- Rejmanek, D., Foley, P., Barbet, A., & Foley, J. (2012). Evolution of antigen variation in the tick-borne pathogen *Anaplasma phagocytophilum*. *Molecular biology and evolution*, 29(1), 391-400.
- Renesto, P., Gautheret, D., Drancourt, M., & Raoult, D. (2000). Determination of the *rpoB* gene sequences of *Bartonella henselae* and *Bartonella quintana* for phylogenetic analysis. *Research in microbiology*, 151(10), 831-836.
- Resto-Ruiz, S., Burgess, A., & Anderson, B. E. (2003). The role of the host immune response in pathogenesis of *Bartonella henselae*. *DNA and cell biology*, 22(6), 431-440.
- Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*, 106(45), 19126-19131.
- Riess, T., Andersson, S. G., Lupas, A., Schaller, M., Schäfer, A., Kyme, P., ... & Kempf, V. A. (2004). *Bartonella* adhesin a mediates a proangiogenic host cell response. *The Journal of experimental medicine*, 200(10), 1267-1278.
- Rikihisa, Y. (2011). Mechanisms of obligatory intracellular infection with *Anaplasma phagocytophilum*. *Clinical microbiology reviews*, 24(3), 469-489.
- Rikihisa, Y. (2017). Role and function of the type IV secretion system in *Anaplasma* and *Ehrlichia* species. *Type IV Secretion in Gram-Negative and Gram-Positive Bacteria*, 297-321.
- Robinson, M. T., Shaw, S. E., & Morgan, E. R. (2009). *Anaplasma phagocytophilum* infection in a multi-species deer community in the New Forest, England. *European Journal of Wildlife Research*, 55, 439-442.
- Rolain, J. M., Franc, M., Davoust, B., & Raoult, D. (2003). Molecular detection of *Bartonella quintana*, *B. koehlerae*, *B. henselae*, *B. clarridgeiae*, *Rickettsia felis*, and *Wolbachia pipientis* in cat fleas, France. *Emerging infectious diseases*, 9(3), 339.
- Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., ... & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome biology*, 14, 1-20.
- Ruiz-Fons, F., & Gilbert, L. (2010). The role of deer as vehicles to move ticks, *Ixodes ricinus*, between contrasting habitats. *International journal for parasitology*, 40(9), 1013-1020.
- Saenz, H. L., Engel, P., Stoeckli, M. C., Lanz, C., Raddatz, G., Vayssier-Taussat, M., ... & Dehio, C. (2007). Genomic analysis of *Bartonella* identifies type IV secretion systems as host adaptability factors. *Nature genetics*, 39(12), 1469-1476.
- Salata, C., Moutailler, S., Attoui, H., Zweygarth, E., Decker, L., & Bell-Sakyi, L. (2021). How relevant are in vitro culture models for study of tick-pathogen interactions?. *Pathogens and global health*, 115(7-8), 437-455.
- Sanchez Clemente, N., Ugarte-Gil, C. A., Solórzano, N., Maguiña, C., Pachas, P., Blazes, D., ... & Moore, D. (2012). *Bartonella bacilliformis*: a systematic review of the literature to guide the research agenda for elimination. *PLoS neglected tropical diseases*, 6(10), e1819.

- Schäfer, I., & Kohn, B. (2020). *Anaplasma phagocytophilum* infection in cats: A literature review to raise clinical awareness. *Journal of feline medicine and surgery*, 22(5), 428-441.
- Scharf, W., Schauer, S., Freyburger, F., Petrovec, M., Schaarschmidt-Kiener, D., Liebisch, G., ... & Von Loewenich, F. D. (2011). Distinct host species correlate with *Anaplasma phagocytophilum* ankA gene clusters. *Journal of clinical microbiology*, 49(3), 790-796.
- Schildkraut, C. L., Marmur, J., & Doty, P. (1961). The formation of hybrid DNA molecules and their use in studies of DNA homologies. *Journal of molecular biology*, 3(5), 595-IN16.
- Schulein, R., Guye, P., Rhomberg, T. A., Schmid, M. C., Schröder, G., Vergunst, A. C., ... & Dehio, C. (2005). A bipartite signal mediates the transfer of type IV secretion substrates of *Bartonella henselae* into human cells. *Proceedings of the National Academy of Sciences*, 102(3), 856-861.
- Scorpio, D. G., Leutenegger, C., Berger, J., Barat, N., Madigan, J. E., & Dumler, J. S. (2008). Sequential analysis of *Anaplasma phagocytophilum* *msp2* transcription in murine and equine models of human granulocytic anaplasmosis. *Clinical and Vaccine Immunology*, 15(3), 418-424.
- Scott, G. R., & Horsburgh, D. (1983). New rickettsial isolates. In *Newsletter No. 36*. Centre for Tropical Veterinary Medicine, Royal (Dick) School of Veterinary Studies Edinburgh.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068-2069.
- Shih, C. M., Chao, L. L., & Yu, C. P. (2002). Chemotactic migration of the Lyme disease spirochete (*Borrelia burgdorferi*) to salivary gland extracts of vector ticks. *The American journal of tropical medicine and hygiene*, 66(5), 616-621.
- Siamer, S., & Dehio, C. (2015). New insights into the role of *Bartonella* effector proteins in pathogenesis. *Current opinion in microbiology*, 23, 80-85.
- Silaghi, C., Kauffmann, M., Passos, L. M., Pfister, K., & Zweggarth, E. (2011). Isolation, propagation and preliminary characterisation of *Anaplasma phagocytophilum* from roe deer (*Capreolus capreolus*) in the tick cell line IDE8. *Ticks and tick-borne diseases*, 2(4), 204-208.
- Silaghi, C., Woll, D., Hamel, D., Pfister, K., Mahling, M., & Pfeffer, M. (2012). *Babesia* spp. and *Anaplasma phagocytophilum* in questing ticks, ticks parasitizing rodents and the parasitized rodents—analyzing the host-pathogen-vector interface in a metropolitan area. *Parasites & vectors*, 5, 1-14.
- Spach, D. H., Liles, W. C., Campbell, G. L., Quick, R. E., Anderson Jr, D. E., & Fritsche, T. R. (1993). Tick-borne diseases in the United States. *New England Journal of Medicine*, 329(13), 936-947.
- Stamatakis, A. (2015). Using RAxML to infer phylogenies. *Current protocols in bioinformatics*, 51(1), 6-14.
- Stuen, S., Bergström, K., Petrovec, M., Van de Pol, I., & Schouls, L. M. (2003). Differences in clinical manifestations and hematological and serological responses after experimental infection with genetic variants of *Anaplasma phagocytophilum* in sheep. *Clinical and Vaccine Immunology*, 10(4), 692-695.
- Stuen, S., Granquist, E. G., & Silaghi, C. (2013). *Anaplasma phagocytophilum*—a widespread multi-host pathogen with highly adaptive strategies. *Frontiers in cellular and infection microbiology*, 3, 31.

- Stuen, S., Granquist, E. G., & Silaghi, C. (2013). *Anaplasma phagocytophilum*—a widespread multi-host pathogen with highly adaptive strategies. *Frontiers in cellular and infection microbiology*, 3, 31.
- Stuen, S., Van De Pol, I., Bergström, K., & Schouls, L. M. (2002). Identification of *Anaplasma phagocytophila* (formerly *Ehrlichia phagocytophila*) variants in blood from sheep in Norway. *Journal of Clinical Microbiology*, 40(9), 3192-3197.
- Takumi, K., Hofmeester, T. R., & Sprong, H. (2021). Red and fallow deer determine the density of *Ixodes ricinus* nymphs containing *Anaplasma phagocytophilum*. *Parasites & Vectors*, 14, 1-9.
- Takumi, K., Sprong, H., & Hofmeester, T. R. (2019). Impact of vertebrate communities on *Ixodes ricinus*-borne disease risk in forest areas. *Parasites & Vectors*, 12, 1-12.
- Tamarit, D., Neuvonen, M. M., Engel, P., Guy, L., & Andersson, S. G. (2018). Origin and evolution of the *Bartonella* gene transfer agent. *Molecular biology and evolution*, 35(2), 451-464.
- Telford, S. R., Stewart, P. E., & Bloom, M. E. (2024). Increasing Risk for Tick-Borne Disease: What Should Clinicians Know?. *JAMA Internal Medicine*.
- Tettelin, H., & Medini, D. (2020). The pangenome: Diversity, dynamics and evolution of genomes.
- Thoendel, M., Jeraldo, P. R., Greenwood-Quaintance, K. E., Yao, J. Z., Chia, N., Hanssen, A. D., ... & Patel, R. (2016). Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *Journal of microbiological methods*, 127, 141-145.
- Tolkacz, K., Alsarraf, M., Kowalec, M., Dwużnik, D., Grzybek, M., Behnke, J. M., & Bajer, A. (2018). *Bartonella* infections in three species of *Microtus*: prevalence and genetic diversity, vertical transmission and the effect of concurrent *Babesia microti* infection on its success. *Parasites & vectors*, 11, 1-15.
- Tolnai, Z., Sréter-Lancz, Z., & Sréter, T. (2015). Spatial distribution of *Anaplasma phagocytophilum* and *Hepatozoon canis* in red foxes (*Vulpes vulpes*) in Hungary. *Ticks and tick-borne diseases*, 6(5), 645-648.
- Tuomi, J. (1967). Experimental studies on bovine tick-borne fever. 1. Clinical and haematological data, some properties of the causative agent, and homologous immunity.
- Vandamme, P. A. (2011). Taxonomy and classification of bacteria. *Manual of clinical microbiology*, 210-227.
- Varma, M. G. R., Pudney, M., & Leake, C. J. (1975). The establishment of three cell lines from the tick *Rhipicephalus appendiculatus* (Agari: ixodidae) and their Infection with some arboviruses. *Journal of medical entomology*, 11(6), 698-706.
- Vayssier-Taussat, M., Le Rhun, D., Deng, H. K., Biville, F., Cescau, S., Danchin, A., ... & Dehio, C. (2010). The Trw type IV secretion system of *Bartonella* mediates host-specific adhesion to erythrocytes. *PLoS pathogens*, 6(6), e1000946.
- Vayssier-Taussat, M., Moutailler, S., Féménia, F., Raymond, P., Croce, O., La Scola, B., ... & Raoult, D. (2016). Identification of novel zoonotic activity of *Bartonella* spp., France. *Emerging Infectious Diseases*, 22(3), 457.
- Vermi, W., Facchetti, F., Riboldi, E., Heine, H., Scutera, S., Stornello, S., ... & Musso, T. (2006). Role of dendritic cell-derived CXCL13 in the pathogenesis of *Bartonella henselae* B-rich granuloma. *Blood*, 107(2), 454-462.

von Loewenich, F. D., Baumgarten, B. U., Schröppel, K., Geißdörfer, W., Röllinghoff, M., & Bogdan, C. (2003). High diversity of ankA sequences of *Anaplasma phagocytophilum* among *Ixodes ricinus* ticks in Germany. *Journal of clinical microbiology*, 41(11), 5033-5040.

Vos, M., & Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *The ISME journal*, 3(2), 199-208.

Wagner, A., & Dehio, C. (2019). Role of distinct type-IV-secretion systems and secreted effector sets in host adaptation by pathogenic *Bartonella* species. *Cellular microbiology*, 21(3), e13004.

Wallden, K., Rivera-Calzada, A., & Waksman, G. (2010). Microreview: Type IV secretion systems: versatility and diversity in function. *Cellular microbiology*, 12(9), 1203-1212.

Wang, F., Yan, M., Liu, A., Chen, T., Luo, L., Li, L., ... & Bao, F. (2020). The seroprevalence of *Anaplasma phagocytophilum* in global human populations: A systematic review and meta-analysis. *Transboundary and Emerging Diseases*, 67(5), 2050-2064.

Wells, M. B., & Andrew, D. J. (2019). Anopheles salivary gland architecture shapes Plasmodium sporozoite availability for transmission. *MBio*, 10(4), 10-1128.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial genomics*, 3(10), e000132.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, 13(6), e1005595.

Wilson, M. R., Naccache, S. N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., ... & Chiu, C. Y. (2014). Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *New England Journal of Medicine*, 370(25), 2408-2417.

Woldehiwet, Z. (1987). The effects of tick-borne fever on some functions of polymorphonuclear cells of sheep. *Journal of comparative pathology*, 97(4), 481-485.

Woldehiwet, Z. (2006). *Anaplasma phagocytophilum* in ruminants in Europe. *Annals of the New York Academy of Sciences*, 1078(1), 446-460.

Woldehiwet, Z., & Horrocks, B. K. (2005). Antigenicity of ovine strains of *Anaplasma phagocytophilum* grown in tick cells and ovine granulocytes. *Journal of comparative pathology*, 132(4), 322-328.

Woldehiwet, Z., Horrocks, B. K., Scaife, H., Ross, G., Munderloh, U. G., Bown, K., ... & Hart, C. A. (2002). Cultivation of an ovine strain of *Ehrlichia phagocytophila* in tick cell cultures. *Journal of comparative pathology*, 127(2-3), 142-149.

Yang, J., Liu, Z., Guan, G., Liu, Q., Li, Y., Chen, Z., ... & Yin, H. (2013). Prevalence of *Anaplasma phagocytophilum* in ruminants, rodents and ticks in Gansu, north-western China. *Journal of medical microbiology*, 62(2), 254-258.

Yigit, E., Feehery, G. R., Langhorst, B. W., Stewart, F. J., Dimalanta, E. T., Pradhan, S., ... & Davis, T. B. (2016). A Microbiome DNA Enrichment Method for Next-Generation Sequencing Sample Preparation. *Current Protocols in Molecular Biology*, 115(1), 7-26.

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek*, 110, 1281-1286.

Yunker, C. E., Cory, J., & Meibos, H. (1981). Continuous cell lines from embryonic tissues of ticks (Acari: Ixodidae). *In vitro*, 17(2), 139-142.

Yunker, C. E., Cory, J., & Meibos, H. (1984). Tick tissue and cell culture: applications to research in medical and veterinary acarology and vector-borne disease. *Acarology VI/editors, DA Griffiths and CE Bowman*.

Zähringer, U., Lindner, B., Knirel, Y. A., van den Akker, W. M., Hiestand, R., Heine, H., & Dehio, C. (2004). Structure and biological activity of the short-chain lipopolysaccharide from *Bartonella henselae* ATCC 49882T. *Journal of Biological Chemistry*, 279(20), 21046-21054.

Zeineldin, M., Camp, P., Farrell, D., Lehman, K., & Thacker, T. (2023). Whole genome sequencing of *Mycobacterium bovis* directly from clinical tissue samples without culture. *Frontiers in Microbiology*, 14, 1141651.

Zhan, L., Cao, W. C., Jiang, J. F., Zhang, X. A., Liu, Y. X., Wu, X. M., ... & Habbema, J. D. F. (2010). *Anaplasma phagocytophilum* from rodents and sheep, China. *Emerging infectious diseases*, 16(5), 764.

Zhang, P., Jiang, D., Wang, Y., Yao, X., Luo, Y., & Yang, Z. (2021). Comparison of de novo assembly strategies for bacterial genomes. *International Journal of Molecular Sciences*, 22(14), 7668



# APPENDIX

## 1.0.0: Sheep spleens metadata

Date Collected	Farm	ID	Host	Postcode	PCR Result	CT
08/11/2022		1 JI01	Sheep	DG14 0YG	N	
08/11/2022		1 JI02	Sheep	DG14 0YG	N	
08/11/2022		1 JI03	Sheep	DG14 0YG	N	
08/11/2022		1 JI04	Sheep	DG14 0YG	N	
08/11/2022		2 DHB01	Sheep	CA7 0NH	N	
08/11/2022		2 DHB02	Sheep	CA7 0NH	N	
08/11/2022		2 DHB03	Sheep	CA7 0NH	N	
08/11/2022		3 AEH01	Sheep	CA19 1TH	N	
08/11/2022		3 AEH02	Sheep	CA19 1TH	N	
08/11/2022		3 AEH03	Sheep	CA19 1TH	P	24.93
08/11/2022		3 AEH04	Sheep	CA19 1TH	N	
08/11/2022		3 AEH05	Sheep	CA19 1TH	N	
08/11/2022		3 AEH06	Sheep	CA19 1TH	N	
08/11/2022		3 AEH07	Sheep	CA19 1TH	P	21.64
08/11/2022		3 AEH08	Sheep	CA19 1TH	N	
08/11/2022		4 PKM	Sheep	LA18 4AR	N	
08/11/2022		5 TAB01	Sheep	LA12 7XA	N	
08/11/2022		5 TAB02	Sheep	LA12 7XA	N	
08/11/2022		5 TAB03	Sheep	LA12 7XA	P	29.95
08/11/2022		5 TAB04	Sheep	LA12 7XA	N	
08/11/2022		5 TAB05	Sheep	LA12 7XA	N	
08/11/2022		5 TAB06	Sheep	LA12 7XA	N	
08/11/2022		5 TAB07	Sheep	LA12 7XA	N	
08/11/2022		5 TAB08	Sheep	LA12 7XA	N	
08/11/2022		6 JEA01	Sheep	LA12 8DB	P	29.99
08/11/2022		6 JEA02	Sheep	LA12 8DB	P	29.88
08/11/2022		6 JEA03	Sheep	LA12 8DB	P	27.7
08/11/2022		7 WR01	Sheep	LA22 6RS	N	
08/11/2022		7 WR02	Sheep	LA22 6RS	N	
08/11/2022		7 WR03	Sheep	LA22 6RS	N	
08/11/2022		7 WR04	Sheep	LA22 6RS	P	27.88
08/11/2022		8 JRW01	Sheep	LA12 7ON	N	
08/11/2022		8 JRW02	Sheep	LA12 7ON	N	
08/11/2022		8 JRW03	Sheep	LA12 7ON	N	
08/11/2022		8 JRW04	Sheep	LA12 7ON	N	
08/11/2022		9 LRH01	Sheep	LA5 9QE	N	
08/11/2022		9 LRH02	Sheep	LA5 9QE	N	
08/11/2022		9 LRH03	Sheep	LA5 9QE	N	
08/11/2022		9 LRH04	Sheep	LA5 9QE	N	
08/11/2022		9 LRH05	Sheep	LA5 9QE	N	
08/11/2022		10 JMW01	Sheep	LA8 9LD	N	
08/11/2022		10 JMW02	Sheep	LA8 9LD	N	
08/11/2022		10 JMW03	Sheep	LA8 9LD	N	
08/11/2022		11 JL01	Lamb	LA6 3NX	N	
08/11/2022		11 JL02	Lamb	LA6 3NX	N	
08/11/2022		11 JL03	Lamb	LA6 3NX	N	
08/11/2022		11 JL04	Lamb	LA6 3NX	N	

08/11/2022	12 JM03	Sheep	LA8 OHX	N	
08/11/2022	12 JM04	Sheep	LA8 OHX	N	
08/11/2022	12 JM05	Sheep	LA8 OHX	N	
08/11/2022	12 JM06	Sheep	LA8 OHX	N	
08/11/2022	12 JM07	Sheep	LA8 OHX	N	
08/11/2022	12 JM01	Sheep	LA8 OHX	N	
08/11/2022	12 JM02	Sheep	LA8 OHX	N	
08/11/2022	13 JEH01	Lamb	LA10 SHW	N	
08/11/2022	13 JEH02	Lamb	LA10 SHW	N	
08/11/2022	14 DS01	Sheep	LA6 2EU	N	
08/11/2022	14 DS02	Sheep	LA6 2EU	N	
08/11/2022	15 SDL01	Sheep	LA6 2QA	N	
08/11/2022	15 SDL02	Sheep	LA6 2QA	N	
08/11/2022	15 SDL03	Sheep	LA6 2QA	N	
08/11/2022	15 SDL04	Sheep	LA6 2QA	N	
08/11/2022	15 SDL05	Sheep	LA6 2QA	N	
08/11/2022	15 SDL06	Sheep	LA6 2QA	N	
08/11/2022	16 JH01	Sheep	LA62RM	N	
08/11/2022	16 JH02	Sheep	LA62RM	N	
08/11/2022	16 JH03	Sheep	LA62RM	N	
08/11/2022	16 JH04	Sheep	LA62RM	N	
08/11/2022	16 JH05	Sheep	LA62RM	N	
08/11/2022	16 JH06	Sheep	LA62RM	N	
08/11/2022	16 JH07	Sheep	LA62RM	N	
08/11/2022	16 JH08	Sheep	LA62RM	N	
08/11/2022	17 IB01	Sheep	LA6 2VJ	N	
08/11/2022	17 IB02	Sheep	LA6 2VJ	N	
08/11/2022	17 IB03	Sheep	LA6 2VJ	N	
08/11/2022	17 IB04	Sheep	LA6 2VJ	N	
08/11/2022	17 IB05	Sheep	LA6 2VJ	N	
08/11/2022	17 IB06	Sheep	LA6 2VJ	P	25.2
08/11/2022	18 PKRW01	Sheep	LA8 0DZ	N	
08/11/2022	18 PKRW02	Sheep	LA8 0DZ	N	

## 1.1.0: Deer spleens metadata

Date Collected	ID	Host	Age	Sex	PCR Result	CT
n/a	GR01	Red	Juvenile	Female	P	23.49
16/11/2022	GR02	Red	Adult	Female	N	
28/11/2022	GR03	Red	Adult	Female	P	23.57
22/11/2022	GR04	Red	Adult	Female	N	
16/11/2022	GR05	Red	Juvenile	Male	P	21.5
28/11/2022	GR06	Red	Juvenile	Male	P	23.8
16/11/2022	GR07	Red	Adult	Female	P	27.66
09/03/2023	GR08	Red	Adult	Female	P	27.83
13/02/2023	GR09	Red	Juvenile	Female	P	22.98
16/01/2023	GR10	Red	Adult	Female	P	26.59
09/03/2023	GR11	Red	Juvenile	Male	P	23.47
09/12/2022	GR12	Red	Adult	Female	P	26.63
22/03/2023	GR13	Red	Juvenile	Female	P	21.15
22/03/2023	GR14	Red	Adult	Female	P	23.89
07/12/2022	GR15	Red	Adult	Male	P	29.97
12/03/2023	GR16	Red	Yearling	Male	P	25.69
09/03/2023	GR17	Red	Adult	Female	P	30.63
06/12/2022	GR18	Red	Yearling	Male	P	23.95
15/02/2023	GR19	Red	Juvenile	Female	P	25.74
22/03/2023	GR20	Red	Adult	Female	P	25.06
24/03/2023	GR21	Red	Adult	Female	P	24.43
24/03/2023	GR22	Red	Adult	Female	P	26.63
23/03/2023	GR23	Red	Juvenile	Female	P	27.98
18/11/2022	GRD01	Roe	Adult	Female	P	29.96
28/11/2022	GRD02	Roe	Adult	Female	P	28.48
04/11/2022	GRD03	Roe	Adult	Female	P	24.26
04/11/2022	GRD04	Roe	Adult	Female	P	29.72
28/11/2022	GRD05	Roe	Adult	Male	P	27.86
01/11/2022	GRD06	Roe	Adult	Female	P	21.84
09/11/2022	GRD07	Roe	Adult	Female	P	24.17
01/11/2022	GRD08	Roe	Juvenile	Male	P	19.03
09/11/2022	GRD09	Roe	Juvenile	Male	P	20.45
21/11/2022	GRD10	Roe	Adult	Female	N	
10/11/2022	GRD11	Roe	Adult	Female	P	25.5
21/11/2022	GRD12	Roe	Adult	Female	P	26.93
30/11/2022	GRD13	Roe	Juvenile	Female	P	26.72
n/a	GRD14	Roe	Juvenile	Female	P	23.53
n/a	GRD15	Roe	Adult	Female	P	28.77
09/11/2022	GRD16	Roe	Adult	Female	P	30
09/11/2022	GRD17	Roe	Juvenile	Female	P	21.19
01/11/2022	GRD18	Roe	Juvenile	Female	P	19.45
01/11/2022	GRD19	Roe	Juvenile	Female	P	22.27
16/11/2022	GRD20	Roe	Yearling	Female	N	
18/11/2022	GRD21	Roe	Juvenile	Female	P	27.91
10/02/2023	GRD22	Roe	Adult	Female	P	28.73
n/a	GRD23	Roe	Adult	Female	P	25.74
15/12/2022	GRD24	Roe	Yearling	Female	P	27.77
07/03/2023	GRD25	Roe	Adult	Female	P	23.08
21/02/2023	GRD26	Roe	Adult	Female	P	28.38

31/01/2023	GRD27	Roe	Juvenile	Male	P	22.69
09/12/2022	GRD28	Roe	Adult	Female	P	26.69
31/01/2023	GRD29	Roe	Adult	Female	P	30
27/01/2023	GRD30	Roe	Juvenile	Female	P	25.97
20/12/2022	GRD31	Roe	Juvenile	Female	P	20.72
27/01/2023	GRD32	Roe	Adult	Female	P	26.65
27/01/2023	GRD33	Roe	Adult	Female	P	29.16
02/02/2023	GRD34	Roe	Yearling	Female	P	28.13
10/02/2023	GRD35	Roe	Juvenile	Female	P	24.77
20/12/2022	GRD36	Roe	Adult	Female	P	18.81
06/03/2023	GRD37	Roe	Adult	Female	P	26.62
n/a	GRD38	Roe	Yearling	Female	P	27.64
06/02/2023	GRD39	Roe	Adult	Male	N	
27/01/2023	GRD40	Roe	Adult	Female	P	26.52
31/01/2023	GRD41	Roe	Adult	Female	P	27.81
10/02/2023	GRD42	Roe	Adult	Female	N	
14/02/2023	GRD43	Roe	Adult	Female	P	29.69
10/02/2023	GRD44	Roe	Juvenile	Female	P	24.99
10/02/2023	GRD45	Roe	Juvenile	Female	P	25.83
31/01/2023	GRD46	Roe	Yearling	Female	P	24.63
07/02/2023	GRD47	Roe	Yearling	Female	P	26.82
17/04/2023	GRD48	Roe	Yearling	Male	P	27.96
07/03/2023	GRD49	Roe	Adult	Female	N	
17/04/2023	GRD50	Roe	Adult	Male	P	29.5

### 1.2.0: Small mammal spleens metadata

ID	Host	Location	PCR Result	CT
CS1	Shrew	Kielder	N	
CS2	Shrew	Kielder	N	
CS3	Shrew	Kielder	N	
CS4	Shrew	Kielder	P	26.5
CS5	Shrew	Kielder	P	23.5
CS6	Shrew	Kielder	N	
CS7	Shrew	Kielder	N	
CS8	Shrew	Kielder	P	25.9
CS9	Shrew	Kielder	N	
CS10	Shrew	Kielder	P	26
CS11	Shrew	Kielder	N	
CS12	Shrew	Kielder	N	
CS13	Shrew	Kielder	N	
FV1	Field Vole	Kielder	N	
FV2	Field Vole	Kielder	N	
FV3	Field Vole	Kielder	N	

## 2.0.0: *Bartonella* Genome Assembly & Annotation

### 2.0.1: Short Read Processing:

```
fastp -i C271_R1.fastq.gz -I C271_R2.fastq.gz -o  
C271_R1_trimmed.fastq.gz -O C271_R2_trimmed.fastq.gz -q 20
```

### 2.0.2: Long Read Processing:

```
Cat *.fastq.gz > C271.fastq.gz
```

```
Porechop -I C271.fastq.gz -o C271_AR.fastq.gz --threads 24
```

```
filtlong --min_length 1000 -keep_percent 90 target_bases 5000000000  
C271_AR.fastq.gz | gzip > C271_AR_Filt.fastq.gz
```

### 2.0.3: Unicycler Assembly

```
unicycler -1 C271_R1_trimmed.fastq.gz -2 C271_R2_trimmed.fastq.gz -l  
C271_AR_Filt.fastq.gz -o C271_Assembly
```

### 2.0.4: Annotation with Prokka:

```
prokka --outdir C271_prokka --prefix C271 --genus Anaplasma  
C271.fasta
```

### 2.0.5: Annotation with ggCaller

```
ls -d 1 $PWD/*.fasta > Bart_Genomes.txt
```

(This should contain all of the genomes from the *Bartonella* genus currently available for best results)

```
ggcaller --refs Bart_Genomes.txt --out Bartonella_ggcaller
```

(This will output a directory titled: *Bartonella\_ggcaller* containing all of the input genomes annotated by ggCaller).

### 2.0.6: Annotation with RASTtk

(RASTtk annotations were completed on the BV-BRC website following their guides: <https://www.bv-brc.org/app/ComprehensiveGenomeAnalysis>).

All following annotations on *Anaplasma phagocytophilum* strains were performed using the same abovementioned commands with appropriate amendments.

## 2.1.0: *Anaplasma phagocytophilum* UK Strains Assembly

### 2.1.1: Read filtering

```
bowtie2-build ISE6_cell_line.fastq.gz ISE6_DB

bowtie2 -x ISE6_DB -1 harris_R1.fastq.gz -2 harris_R2.fastq.gz -S
harris_mapped_unmapped.sam

samtools view -bS harris_mapped_unmapped.sam >
harris_mapped_unmapped.bam

samtools view -b -f 12 -F 256 harris_mapped_unmapped.bam >
harris_both_unmapped.bam

samtools sort -n harris_both_unmapped.bam > harris_sorted.bam

bedtools bamtofastq -i harris_sorted.bam -fq
harris_host_rem_R1.fastq.gz fq2 harris_host_rem_R2.fastq.gz
```

### 2.1.2: Short Read Processing

```
fastp -i harris_host_rem_R1.fastq.gz -I harris_host_rem_R2.fastq.gz -
o harris_host_rem_R1_trimmed.fastq.gz -O
harris_host_rem_R2_trimmed.fastq.gz -q 20
```

### 2.1.3: Unicycler

```
unicycler -1 harris_host_rem_R1_trimmed.fastq.gz -2
harris_host_rem_R2_trimmed.fastq.gz -l harris_raw.fastq.gz -o
harris_Assembly
```

## 2.2.0 *Anaplasma phagocytophilum* Agilent SureSelect Data Processing

### 2.2.1: Agilent Java Proprietary Trimmer:

```
java -jar
/Users/erg408/Desktop/SureSelect/AGeNT_3.0.6/agent/lib/trimmer-
3.0.5.jar -v2 -fq1 /Users/erg408/Desktop/SureSelect/ -fq2
/Users/erg408/Desktop/SureSelect/ -out
/Users/erg408/Desktop/SureSelect/S5_trimmed
```

### 2.2.2: Kraken 2 Identification of Short Reads

```
kraken2 --use-names --db /home/seanb/SOFTWARE/kraken2/kraken2-
microbial-fatfree --threads 24 --report S5_3.report --paired --
minimum-hit-groups 3 --confidence 0.05 S5_Bacteria_extracted_R1.fastq
S5_Bacteria_extracted_R2.fastq > S5_3.kraken
```

### 2.2.3: Kraken 2 Identification of Long Reads

```
kraken2 --use-names --db /home/seanb/SOFTWARE/kraken2/kraken2-
microbial-fatfree --threads 24 --report S1_pion.report --minimum-hit-
groups 3 --confidence 0.05 S1_trimmed_long.fastq.gz > S1_pion.kraken
```

#### 2.2.4: Extracting Reads from Taxa

```
python3 extract_kraken_reads.py -k S5.kraken -s1
S5_host_removed_R1.fastq -s2 S5_host_removed_R2.fastq -o
S5_Ap_extracted_R1.fastq -o2 S5_Ap_extracted_R2.fastq --taxid 948 --
include-children --fastq-output --report S5.report
```

(Script: `extract_kraken_reads.py` GitHub: [jenniferlu717/KrakenTools](https://github.com/jenniferlu717/KrakenTools))

#### 2.2.5: Mapping Short Reads

```
minimap2 -ax sr Harris.mmi GRD08_R1.fastq GRD08_R2.fastq >
GRD08_short.sam
```

#### 2.2.6: Mapping Long Reads

```
minimap2 -ax map-ont -t 24 Harris.mmi GRD08_all_long.fastq.gz >
GRD08_long.sam
```

#### 2.2.7: Merging Long & Short Read Data

```
samtools view -h GRD08_long.bam GRD08_long.sam
```

```
samtools sort GRD08_long.bam > GRD08_long_sorted.bam
```

(Repeat for short reads).

```
samtools merge GRD08_all.bam GRD08_short_sorted.bam
GRD08_long_sorted.bam
```

#### 2.2.8: Generating Coverage Plot: Bam2plot

```
bam2plot --bam GRD08_all.bam --outpath GRD08_plots --threshold 30 -s
-hl -p svg
```

#### 2.2.9: Generating Gaps Plot: GapPlotter (written for this study)

GapPlotter is a Python package for analyzing and visualizing gaps in DNA sequence alignments stored in BAM format. It provides two scripts:

1. `gapplotter_generate_csv.py`: This script generates a CSV file containing information about gaps in a BAM file. It identifies regions of the reference genome that are not covered by any read and records them as gaps, including their start and end positions.
2. `gapplotter_generate_histogram.py`: This script generates a histogram image of gap lengths from a CSV file generated by GapPlotter. It plots the distribution of gap

lengths, allowing users to visualize the frequency and distribution of gaps in the alignment data.

GapPlotter is useful for researchers working with DNA sequence data who need to analyze and visualize regions of the genome that are not covered by reads, such as gaps in sequence coverage or alignment errors.

## Features

- Generate CSV files of gap information from BAM files
- Visualize gap lengths using histogram plots
- Command-line interface for easy usage
- Compatible with Python 3.x

## Installation

To install GapPlotter, simply clone the repository and install dependencies.

### Installation Instructions

1. Clone the Repository Clone the GapPlotter repository to your local machine using the following command:

```
git clone https://github.com/Wizical/GapPlotter.git
```

2. Install Dependencies Navigate to the cloned repository directory and install the required dependencies using pip

## Usage

**Generating CSV File of Gaps** To generate a CSV file containing information about gaps in a BAM file, use the following command:

```
python gapplotter_generate_csv.py -i input.bam -o output.csv
```

Replace input.bam with the path to your BAM file and output.csv with the desired path and filename for the CSV file.

**Generating Histogram Image** To generate a histogram image of gap lengths from a CSV file generated by GapPlotter, use the following command:

```
python gapplotter_generate_histogram.py -i input.csv -o output.png
```



Replace input.csv with the path to the CSV file containing gap information and output.png with the desired path and filename for the PNG image of the histogram.

### 3.0.0: Annotation Methods Metadata (Chapters 2/3)

**Table 7: Total number of CDS annotated on Seven UK derived isolates of *A. phagocytophilum* (Harris, Perth, ZW144, ZW122, ZW129, OS, FG) and three strains of *Bartonella bennettii* sp. nov. (C271, J117, D105) and *B. heixiaziensis* (RE21).**

Strain	Prokka	RASTtk	ggCaller
C271	1420	1492	1397
J117	1470	1506	1409
D105	1502	1532	1455
RE21	2112	2124	2067
Harris	1681	1723	1260
Perth	1360	1420	1252
ZW144	1325	1378	1246
ZW122	1370	1409	1296
ZW129	1341	1382	1239
OS	1395	1346	1302
FG	1335	1705	1112

### 3.1.0: Assembly Metrics/Sequencing Methods (Chapter 3)

**Table 8: Assembly statistics & sequencing data generated for seven UK derived isolates of *A. phagocytophilum*.**

Strain	Contigs	Size	CDS	Method
Harris	1	1,560,625	1723	ONT, (Illumina Outsourced)
Perth	55	1,544,003	1420	ONT, (Illumina Outsourced)
OS	1	1,609,179	1346	ONT, Illumina (By University of Liverpool)
FG	186	1,520,036	1705	ONT, Illumina v2
ZW122	6	1,554,927	1409	ONT, (Illumina Outsourced)
ZW129	1	1,518,219	1382	ONT, (Illumina Outsourced)
ZW144	38	1,515,251	1378	ONT, (Illumina Outsourced)