Intra-specific variation in avian malaria

Linking infection dynamics to haplotypes

Kalbskopf, Victor

2023

Document Version:
Publisher's PDF, also known as Version of record

Intra-specific variation in avian malaria

Linking infection dynamics to haplotypes

VICTOR KALBSKOPF
DEPARTMENT OF BIOLOGY | FACULTY OF SCIENCE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

Faculty of Science
Department of Biology

ISBN 978-91-8039-771-1

Intra-specific variation in avian malaria

Linking infection dynamics to haplotypes

Victor Kalbskopf

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Science at Lund University to be publicly defended on date of 22 September at 09.00 in Blue Hall, Ecology Building, Department of Biology, Sölvegatan 37, Lund, Sweden

Faculty opponent
Professor Ravinder Sehgal
Department of Biology
San Francisco State University

Organization: LUND UNIVERSITY

Document name: Doctoral dissertation **Date of issue** 2023-08-29

Author(s): Victor Kalbskopf **Sponsoring organization:**

Title and subtitle: Intra-specific variation in avian malaria: Linking infection dynamics to haplotypes

Abstract:

Avian Malaria infects thousands of species of birds across the Aves class. The most widespread (geographically and phylogenetically) morphological species is *Plasmodium relictum,* and of that, the mitochondrial lineage of SGS1 is the most common and invasive. It infects 146 bird species, with large differences in infection outcome, from low intensity chronic disease to shorter, highly virulent infections. In Paper 1, I studied the gene expression of SGS1 in experimentally infected birds pre and post maximum parasitemia. The parasitemia varied widely between birds and especially over the course of the infection. These differences were significantly linked to genes related to cell replication and cellular movement in high parasitemia infections and cellular metabolism in low parasitemia infections. We found that over time, variation in gene expression increased between samples, possibly illustrating individual responses of the parasites to their hosts, and a desynchronisation in their lifecycles. Paper 2 explored the phylogeography of SGS1, and its related lineage GRW11, in the palearctic region. Because the lineage system is defined by a highly conserved single mitochondrial gene, the nuclear polymorphic cell invasion gene, merozoite surface protein 1 (*msp1*), was selected to study the genetic variation present in infected resident and migrant host populations. We found extremely little variation, suggesting SGS1and GRW11 in Europe have an epidemic population structure, or there is strong purifying selection pressure on the *msp1* gene despite the wide host range. Paper 3 developed a genomic sequence capture method using 1035 probes designed for SGS1, and tested it on a range of SGS1, GRW11, and GRW4 samples. The probes effectively isolated DNA from all three lineages, but sequencing success was low for samples with less than 1% parasitemia. We selected 25 genes to describe the higher-than-expected variation within SGS1 and with GRW11 and GRW4 samples. In Paper 4, two different host sources of SGS1 infected blood were used to infect two groups of canaries. The groups differed in parasitemia and mortality, and from each group the three birds with largest differences in infection outcome were selected for RNA sequencing to survey the underlying genomic variation. The source of the infection reliably separated the samples phylogenetically, with relatively less variation observed within the groups. This suggests that an infection is made up of a population of genetically diverse parasites. Paper 5 expanded on this idea by using the genomic sequence capture method from Paper 3 and refined bioinformatic methods from Paper 4 on some of the same samples from Paper 1. Samples collected at the same time points (8 and 20 days post infection) were sequenced. This allowed analyses of how the predominant haplotypes change during an infection, and then link those haplotypes to the disease severity. We found that the least suppressed/most virulent haplotypes had genetic variants in genes related to cell invasion and immune evasion. The combined results of my thesis have far-reaching implications that extend beyond the particular organism under investigation. The notion of genetic diversity within a single infections and the resulting parasite population dynamics offers exciting prospects for future research.

Key words: Avian Malaria, RNA-seq, targeted sequencing, population genetics, parasite evolution, host-parasite interactions, phylogenetics, phylogeography

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language English **ISSN and key title**:

ISBN: 978-91-8039-772-8 (electronic)
978-91-8039-771-1 (print)

Recipient's notes **Number of pages**: 80

Price Security classification

Signature Date 2023-08-04

Intra-specific variation in avian malaria

Linking infection dynamics to haplotypes

Victor Kalbskopf

LUND
UNIVERSITY

Cover Illustration by Victor Kalbskopf

Faculty of Science

Department of Biology

ISBN 978-91-8039-772-8 (electronic)
978-91-8039-771-1 (print)

Printed in Sweden by Media-Tryck, Lund University

Lund 2022

To Synne, my wife, my supporter. You don't know what you mean to me.

Table of Contents

1.1 Abstract

Avian Malaria infects thousands of species of birds across the Aves class. The most widespread (geographically and phylogenetically) morphological species is *Plasmodium relictum*, and of that, the mitochondrial lineage of SGS1 is the most common and invasive. It infects 146 bird species, with large differences in infection outcome, from low intensity chronic disease to shorter, highly virulent infections.

In Paper 1, I studied the gene expression of SGS1 in experimentally infected birds pre and post maximum parasitemia. The parasitemia varied widely between birds and especially over the course of the infection. These differences were significantly linked to genes related to cell replication and cellular movement in high parasitemia infections and cellular metabolism in low parasitemia infections. We found that over time, variation in gene expression increased between samples, possibly illustrating individual responses of the parasites to their hosts, and a desynchronisation in their lifecycles.

Paper 2 explored the phylogeography of SGS1, and its related lineage GRW11, in the palearctic region. Because the lineage system is defined by a highly conserved single mitochondrial gene, the nuclear polymorphic cell invasion gene, merozoite surface protein 1 (msp1), was selected to study the genetic variation present in infected resident and migrant host populations. We found extremely little variation, suggesting SGS1and GRW11 in Europe have an epidemic population structure, or there is strong purifying selection pressure on the msp1 gene despite the wide host range.

Paper 3 developed a genomic sequence capture method using 1035 probes designed for SGS1, and tested it on a range of SGS1, GRW11, and GRW4 samples. The probes effectively isolated DNA from all three lineages, but sequencing success was low for samples with less than 1% parasitemia. We selected 25 genes to describe the higher-than-expected variation within SGS1 and with GRW11 and GRW4 samples.

In Paper 4, two different host sources of SGS1 infected blood were used to infect two groups of canaries. The groups differed in parasitemia and mortality, and from each group the three birds with largest differences in infection outcome were selected for RNA sequencing to survey the underlying genomic variation. The source of the infection reliably separated the samples phylogenetically, with relatively less variation observed within the groups.

This suggests that an infection is made up of a population of genetically diverse parasites.

Paper 5 expanded on this idea by using the genomic sequence capture method from Paper 3 and refined bioinformatic methods from Paper 4 on some of the same samples from Paper 1. Samples collected at the same time points (8 and 20 days post infection) were sequenced. This allowed analyses of how the predominant haplotypes change during an infection, and then link those haplotypes to the disease severity. We found that the least suppressed/most virulent haplotypes had genetic variants in genes related to cell invasion and immune evasion.

The combined results of my thesis have far-reaching implications that extend beyond the particular organism under investigation. The notion of genetic diversity within a single infections and the resulting parasite population dynamics offers exciting prospects for future research.

1.2 Popular Science Summary

"Mal aria" was coined by the ancient Romans, and directly translated from Latin means "bad air", as it was thought that the swampy, humid atmosphere around Italian marshes was responsible for the feverish diseases they would suffer from. It was thousands of years later in 1880 that it was discovered that the disease Malaria was caused by a microscopic parasite which forces its way into red blood cells where they reproduce many times. Thanks to experiments in birds, Ronald Ross realised that bird malaria was transmitted by mosquitos. This was later confirmed for human malaria too. Bird malaria is very similar to human malaria, and in fact we can find malaria-like parasites infecting many mammals and reptiles. In humans, malaria is caused by five species of *Plasmodium* (like *Plasmodium falciparum*), but in birds it's caused by many more species of *Plasmodium* (and other groups like *Plasmodium*). Bird malaria is found on every continent except Antarctica, even in cold climates like Northern Europe, but thankfully it can't infect humans because bird red blood cells are extremely different from human red blood cells. However, birds make it much easier to study malaria, because we can do experiments on them which we can't do on humans.

There is one species of bird malaria called *P. relictum* which infects the greatest number of bird species, and sometimes it's very deadly or very mild, even when infecting the same bird species. We thought that it mostly depended on how strong the bird was that determined whether it died from an infection, but it turns out to be more complicated. I developed methods to find and compare differences in the genes of the parasite between infections, because understanding the genetic variations among the parasites could potentially shed light on how *P. relictum* can infect so many species and why its disease can vary so much when infecting the same bird species.
When a bird is infected, many parasites enter the blood, but we discovered that they are not all identical, and each parasite will use slightly different strategies to evade the bird's immune system and reproduce. This means that when a group of parasites enters a bird during a mosquito bite, a different group of parasites will be transmitted to the next bird, which gives rise to change and evolution of the malaria in just one cycle of infections. We were surprised to see so many differences in the genes in infections and how quickly some parasites took over, possibly causing severe infections in the birds. On the other hand, we noticed that in some birds, the group of parasites didn't change much over time.

1.3 Populärvetenskaplig sammanfattning

"Malaria" myntades av de gamla romarna, och direkt översatt från latin betyder det "dålig luft", för man trodde att den sumpiga, fuktiga atmosfären runt italienska myrar var anledningen för febersjukdomar de drabbades av. Tusentals år senare i 1880 upptäckte man att sjukdomen malaria orsakades av en mikroskopisk parasit som tränger sig in i röda blodkroppar där de förökar sig. Efter Ronald Ross fågelexperiment, visade det sig att fågelmalaria överfördes av myggor. Detta bekräftades senare även för mänsklig malaria. Fågelmalaria är mycket lik mänsklig malaria, och vi hittar malarialiknande parasiter som infekterar många däggdjur och reptiler. Hos människor orsakas malaria av fem arter av *Plasmodium* (som *Plasmodium falciparum*), men hos fåglar orsakas det av många fler arter av *Plasmodium* (och andra grupper som liknar *Plasmodium*). Fågelmalaria finns på alla kontinenter utom Antarktis, även i kalla klimat i norra Europa, men den infekterar inte människor, då fåglars röda blodkroppar är väldigt olika från människors röda blodkroppar. Även med olikheterna är fåglar nyttiga för att studera malaria, då experiment som är omöjliga med människor kan utföras med fåglar.

Den art fågelmalaria som infekterar flest fågelarter heter *P. relictum*. Den kan variera från mild till dödlig, även i samma fågelart. Det troddes att variansen berodde på fågelns kvalitet, men det visar sig vara mer komplicerat än så. Jag utvecklade metoder för att hitta och jämföra skillnader i parasiternas gener mellan infektioner, då förståelsen av den genetiska variationen mellan parasiterna kan hjälpa oss förstå hur *P. relictum* kan smitta så många olika arter samt varför sjukdomen kan variera så mycket även inom samma fågelart.

När en fågel infekteras kommer ett stort antal parasiter in i blodet, men vi upptäckte att alla inte är identiska, och varje parasit använder olika strategier för att undvika fågelns immunsystem och föröka sig. Detta innebär att när en grupp parasiter kommer in i en fågel under ett myggbett, kommer en annan grupp parasiter att överföras till nästa fågel, vilket ger upphov till förändring och utveckling av malaria inom en enda cykel av infektioner. Vi förvånades över att se många genetiska skillnader i infektioner och hur snabbt vissa parasiter dominerade, vilket kan ha orsakat allvarliga infektioner hos fåglarna. Å andra sidan märkte vi även att hos vissa fåglar förändrades knappt gruppen parasiter alls. Detta betyder att vi måste vara mer uppmärksamma på vad som utgör infektioner, då variationen inom gruppen påverkar allvarligheten av sjukdomen.

1.4 List of Papers

Paper 1

Victor Kalbskopf, Dag Ahrén, Gediminas Valki¯unas, Vaidas Palinauskas, Olof Hellgren. 2021. Shifts in gene expression variability in the blood-stage of *Plasmodium relictum*. *Gene,* 792, doi:10.1016/j.gene.2021.145723

Paper 2

Olof Hellgren, **Victor Kalbskopf,** Vincenzo A. Ellis, Arif Ciloglu, Mélanie Duc, Xi Huang, Ricardo J. Lopes, Vanessa A. Mata, Sargis A. Aghayan, Abdullah Inci, Sergei V. Drovetski. Low MSP-1 haplotype diversity in the West Palearctic population of the avian malaria parasite *Plasmodium relictum*. 2021. *Malaria Journal*, 20, no 265, doi:10.1186/s12936-021-03799-8

Paper 3

Vincenzo A. Ellis, **Victor Kalbskopf,** Arif Ciloglu, Mélanie Duc, Xi Huang, Abdullah Inci, Staffan Bensch, Olof Hellgren, Vaidas Palinauskas. Genomic sequence capture of *Plasmodium relictum* in experimentally infected birds. 2022. *Parasites Vectors* 15, 267. doi:10.1186/s13071-022-05373-w

Paper 4

Victor Kalbskopf, Justė Aželytė, Vaidas Palinauskas and Olof Hellgren. 2023. Linking infection outcome to genomic variation within the avian malaria lineage SGS1, *Plasmodium relictum* using experimental infections and RNA - sequencing. Manuscript.

Paper 5

Victor Kalbskopf, Vincenzo Ellis, Dag Ahrén, Vaidas Palinauskas and Olof Hellgren. 2023. Shifts in within-infection haplotype frequencies over the course of single Avian malaria infections and possible links to parasitemia. Manuscript.

1.5 Author's contribution to the papers

Paper 1

VK designed the study with input from OH. GV, VP designed and performed the experiment and provided input to the manuscript. **VK** performed the bioinformatics with supervision from DA and wrote the manuscript. All authors read and approved the final manuscript.

Paper 2

OH designed the study and conducted molecular screening of MSP1. OH and SVD interpreted the results and wrote the first draft. **VK** was responsible for all the bioinformatics, VAE together with AC, MD and XH developed and conducted all the sequence capture laboratory work. SVD, RJL, VAM were responsible for the field sampling of the majority of the used samples. VAM and SAA conducted initial screening of large sample sets to identify those with single SGS1 or GRW11 infections from Portugal/Morocco and Armenia/Russia, respectively. All authors read, commented and approved the final manuscript.

Paper 3

OH and VP designed the study. OH designed the sequence capture probe-set and VP conducted experimental infections and measurements of parasitemia. VAE, AC, MD, XH conducted the sequence capture laboratory work. **VK** developed and conducted all the bioinformatics and statistical analyses. VAE wrote the manuscript with input from all authors. All authors read, commented on and approved the final manuscript.

Paper 4

VK, VP designed the study with input from OH. VP performed the infection experiment. JA prepared figures, statistical analysis, described the methods and results regarding the infection experiment. **VK** developed the bioinformatic pipeline and performed all the analyses related to the transcriptomes. **VK** wrote the manuscript with input from the other authors. All authors read and approved the final manuscript.

Paper 5

VK designed the study with input from OH, DA provided bioinformatics support, VE provided the sequence capture material. **VK** developed the

bioinformatic pipeline and performed the analysis. VP conducted experimental infections and measurements of parasitemia. **VK** wrote the manuscript with input from the other authors. All authors read and approved the final manuscript.

1.6 Abbreviations

bp	base pairs
Mbp	Mega base pairs (1x106)
Gbp	Giga base pairs (1x109)
dpi	days post infection
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid

2 Background

2.1 Introduction

“Successful systems attract parasites” (Despommier, 2016) is an often quoted saying that applies to all of life, from the level of viroids to human corporations. Any self-organising system (organism) that takes resources from another organism (host) to reproduce while causing damage to it (virulence) can be defined as a parasite. Parasites are the only cause of communicable diseases (as opposed to diseases caused by genetics and inappropriate environmental exposure), therefor understanding how they spread and cause disease is both interesting and important as they are a ubiquitous component of ecosystems, and their interactions with hosts and other organisms can have far-reaching effects on the structure and function of ecosystems.

Parasites can impose extremely high levels of selection on their hosts, and the parasites in turn are selected against. They influence the lifecycle, life history traits andsexual selection (Aavani and Rice, 2022) of their hosts and their ecosystem (Kuris *et al.*, 2008). However, the parasites have multiple evolutionary constraints on their success. Higher virulence may result in greater transmission, but may also cause extinction of the host, which is the end of the parasite. Increased host specificity may improve fitness in one host species (due to specialisation in resource utilisation, host manipulation, and immune evasion/suppression), but restrict from infection of related host species and will be bound to the fate of that specific host species. The transmission strategy constrains the evolution of the parasite. For example, sexually transmitted parasites will naturally become host specialist, but parasites associated with environmental transmission have many more opportunities for host generalism. Vectors such as biting insects provide entry past physical host barriers and increase geographical spread, but necessitate additional steps and vector coevolution required to complete its lifecycle.

These evolutionary constraints impose selection, and selection can only act when there is genetic variation in the population. Because parasites inflict high costs on human society and wildlife, understanding the direction and mechanisms of their evolution can provide insights to reduce their impact. They also provide insights into evolution and antagonistic coevolution in the presence conflicting trade-offs. The parasite must spread to anew host and the

host must survive and be able to reproduce, an interaction that is the arms-race of life.

We observe very large variation in disease severity (Gupta *et al.*, 2020) and host specificity (Little, Watt and Ebert, 2006), likely because the outcome of an infection is always due to interactions between the parasite, the host, and the environment (Fecchio, Wells, Bell, V. V. Tkach, *et al.*, 2019), from the level of protein-protein interactions to tectonic plate shifts and ice-ages, all of which varies. To study parasite evolution, the way it adapts to its host, we can leverage that phenotypic variation, and the genetic variation underlying it.

2.2 Parasites

2.2.1 **Transmission**

Parasites transmit to new hosts in a variety of ways. Vertical transmission occurs when offspring acquire the parasite directly from their parent, before, during, or after birth. Environmental transmission frequently requires the faecal-oral route, but sometimes an intermediate host is consumed to complete the parasite's life cycle. Many bacteria and viruses are transmitted directly, where droplets from the respiratory system, skin to skin, or sexual contact is required. Finally, vectors (frequently insects) provide transmission, especially for eukaryotic parasites.

The successful parasite must reproduce within the host to the degree that its descendants will be transmitted to another host. The greater the reproduction within a host, the better its chances of transmission, but this generally increases host mortality (van Baalen and Sabelis, 1995). The virulence/transmission trade-off theory predicts that virulence will be higher in parasites that persist in the environment and do not require direct transmission (Ewald, 1983). The reasoning is that transmission can occur after the death of the host (Ewald, 1983). *Salmonella* and *Vibrio cholerae* are examples of waterborne parasites that show higher case mortality rates in cases of environmental transmission (Ewald, 1994) when compared to pathogens that require direct contact. Walther and Ewald (Walther and Ewald, 2004) found in their meta-analysis of airborne respiratory diseases a strong correlation between virulence and long-lasting propagules in the environment for *Mycobacterium tuberculosis* and the smallpox virus.

Vertical transmission should select for low virulence to ensure transmission of the offspring (Herre, 1993). However, strict vertical transmission is rare, due to the risk of extinction caused by factors (loss of habitat, predation, disease) other than the parasite (Lipsitch *et al.*, 1995). That risk can be reduced if the parasite immunises the host against a horizontally transmitted parasite (Haine, 2008).

Vectors should provide a similar respite from the virulence/transmission trade-off in the same way environment persistence does (Ewald, 1983) since the vector provides a delay in transmission after the death of the host. However, neither the formalised theoretical models nor empirical evidence provides a consistent indication whether or not vector born parasites are more virulent (Cressler *et al.*, 2016). However, since the vectors themselves are parasitised, and they generally have shorter lifespans than the hosts, one would expect that virulence in the vector should be low.

The transmission and virulence of parasites can differ among and within species, with both traits being influenced by the genetic variability they possess. This variability can occur within individual infections as well as across the entire population. In this thesis, my emphasis is on a particular group of parasites that share similar transmission methods, specifically being transmitted by an insect vector during blood feeding. However, there are noticeable variations in virulence among the different species within this group. These blood parasites are commonly referred to as malaria.

2.2.2 Malaria

"Malaria may have killed half of all the people that ever lived" (Whitfield, 2002) is a frequent claim that is very hard to prove but is probably close to the truth (Roser M, 2022). In 2020, WHO estimates that 77% of all malaria related deaths were in children younger than 5 years old and kills about 460 000 children annually (*Malaria deaths by age, World, 1990 to 2019*, no date). The discovery of the causative agent of malaria was made in 1880 (Laveran, 1880), and the vector was discovered thanks to the use avian malaria by Ronald Ross (Ross, 1902). The fact that one could perform experimental infections on birds with infected mosquitos, and not rely on human volunteers, greatly accelerated the progress made in understanding malaria in general. It was only with the development of rodent models and human malaria culturing techniques did avian malaria fall in interest in the research community (Rivero and Gandon, 2018). Nevertheless, while human malaria parasites are widely recognised,

they represent only a small fraction of a vast assemblage of parasites. Within this group, known as Apicomplexa parasites, there are numerous species that share similar life cycles and characteristics in their life histories.

2.2.3 Apicomplexa

The Apicomplexa (also called Apicomplexia) are a large phylum of parasitic alveolates (Levine, 1988). They are a diverse group that includes organisms such as the coccidia, gregarines, piroplasms, haemogregarines, and plasmodia. Diseases caused by Apicomplexa include Babesiosis (*Babesia*), Malaria (*Plasmodium*), Cryptosporidiosis (*Cryptosporidium parvum*), Cyclosporiasis (*Cyclospora cayetanensis*), Cystoisosporiasis (*Cystoisospora belli*), and Toxoplasmosis (*Toxoplasma gondii*) (LEVINE, 1988). Most of them possess a unique form of organelle that comprises a type of non-photosynthetic plastid called an apicoplast. The organelle is an adaptation that the apicomplexan applies in penetration of a host cell (Lim and McFadden, 2010). The groups that infect birds are spread across three families in the order of Haemosporida (Figure 2-1): Haemoproteidae, Leucocytozoidae, and Plasmodiidae.

Haemoproteidae produce pigments, and the two subgenera that infect birds, *Haemoproteus* transmitted by louse flies and Parahaemoproteus transmitted by biting midges, do not have an endo-erythrocytic lifecycle. *Leucocytozoon*, the largest genus of Leucocytozoidae, are primarily transmitted by blackflies (*Simulium*) and are made up of over 100 species. They are a major cause of poultry mortality if their vectors are not controlled. The genera *Plasmodium* of Plasmodiidae are transmitted by mosquitos and cause the titular malaria. The closest thing we have to a type species in avian malaria is *Plasmodium relictum*, the most widespread and common species found to infect birds. *P. relictum* is addressed in detail in the Study species section below.

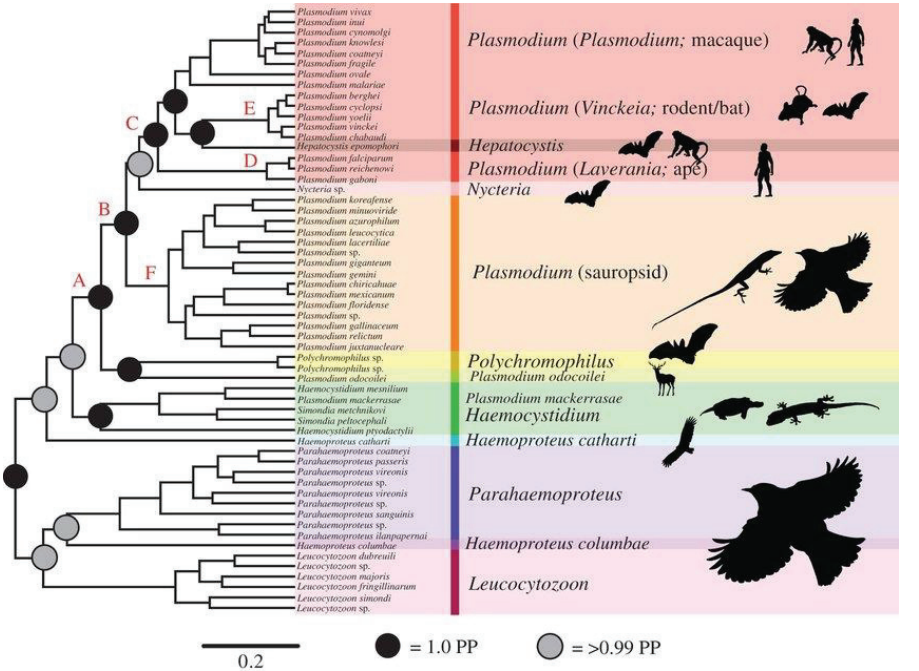

Figure 2-1: Haemosporidian phylogeny based on 21 nuclear genes aligned by amino acids to maintain the reading frame. It includes 58 taxa from 7 genera, which is a small subset of all the lineage that have been found. The colours and silhouettes represent the host groups for those clades. Extracted and modified from Galen et al 2018.

2.3 Avian Malaria

2.3.1 **Lifecycle**

As is the case in human malaria, the lifecycle of avian malaria 'starts' with the mosquito bite.

1. Sporogony in the mosquito: When a mosquito bites an infected bird, it ingests blood containing male and female gametocytes (microgametes and macrogametes, respectively). Inside the mosquito midgut, the microgamete fuses with the macrogamete form a zygote, which is the only diploid stage of the parasite. Each zygote later form

an oocyst in the wall of the midgut of the mosquito where the parasite undergoes multiple steps of meiosis to form sporozoites. The sporozoites migrate to the salivary glands of the mosquito and are ready to infect the next host.

2. Exo-erythrocytic schizogony: During this stage, the sporozoites undergo a process called exo-erythrocytic schizogony, in which they divide into numerous merozoites within the reticuloendothelial cells of the spleen, liver, bone marrow and other immune-related organs.
3. Endo-erythrocytic schizogony: Once the merozoites are produced, they are released into the bloodstream and invade the bird's red blood cells (RBCs). Within the RBCs, the merozoites multiply asexually by schizogony, which causes the host cells to rupture and release more merozoites into the bloodstream. In avian malaria, merozoites can enter the exo-erythrocytic cycle again.
4. A small percentage of the merozoites become gametocytes which are the infective cells which the mosquito takes up.

Each stage of these stages in the malaria lifecycle involves a decrease in the parasite count followed by a process of asexual reproduction. During these events, the parasite population within the vector or the vertebrate host can undergo different levels of bottlenecks, which have implications for the genetic diversity observed in parasite populations. This variability may have an influence on how the host interacts with the parasite.

2.3.2 Malaria as a disease

The main drivers of disease during a malaria infection are the parasite load, the host immune response, and the host genetic factors. The parasite load determines the severity and duration of the infection, as well as the transmission potential to mosquitoes. The host immune response can either control or exacerbate the infection, depending on the type and timing of immune cells and molecules involved (Camponovo *et al.*, 2021). The host genetic factors can influence the susceptibility and resistance to malaria, such as sickle cell trait, glucose-6-phosphate dehydrogenase deficiency, and the Duffy blood group (Phillips *et al.*, 2017) in humans. The typical symptoms of an infection include fever, chills, headache, fatigue, anemia and jaundice (Cowman *et al.*, 2016). Malaria can also lead to complications such as cerebral

malaria, organ failure, low blood sugar, and premature delivery (Phillips *et al.*, 2017; Chua *et al.*, 2021).

The cycles of endo-erythrocytic replication can be synchronised, where blood cells are ruptured simultaneously resulting in periods of fever and other symptoms which occur in cycles (Su, Ruan and Wei, 2011). These periodic symptoms were observed in humans and were recognised as a particular signature of *Plasmodium malariae* (Garcia, Markus and Madeira, 2001), and earlier records of the periodicity of avian malaria indicated that the majority of species exhibited 24-hour cycles or their multiples, and that the peaks could manifest at varying times during the day (Wolfson, 1936; Hewitt, 1940).

The exo-erythrocytic cycle explains another puzzling feature of malaria infections, that of relapse. After the typical symptoms end and the blood smears are clear for a few weeks to months, the patient may relapse. It was demonstrated in avian malaria to be parasites harbouring in the spleen, which can then restart the full infection cycle (Ben-harel, 1923). Relapse can be compared to chronic infections, where there is a low level of parasitemia in blood for periods of years (Hewitt, 1940). The presence of a specific strain will prevent a new infection of the same strain, implying either a form of natural vaccination or competition from coinfection.

The effects of chronic infections in avian malaria are less clear. Knowles *et al.* (2010) showed that low intensity infections in the parents reduced clutch size and offspring health in wild Blue tit populations. Additionally, chronic infections reduce the telomere lengths in migrating great reed warblers, clearly demonstrating long term physiological stress and aging caused by avian malaria (Asghar *et al.*, 2015). A particularly counterintuitive result in experimentally infected birds revealed mortality well after peak parasitemia intensity, apparently caused by blockage of capillaries in the brain leading to cerebral ischaemia (Ilgunas *et al.*, 2016).

It is important to note that individual hosts often exhibit variation in the magnitude of these effects, and the underlying factors contributing to this variation remain unknown. It is unclear whether this variability arises from genetic differences in the host, the parasite, or a combination of both.

2.4 Life-history traits of the parasite and potential roles of genetic variability

While the impact of a disease on its host is often considered the primary focus of study, it is essential to recognise the intricate interplay of various life history traits and interactions. Acquiring a deeper understanding of the intra-genetic variability of parasites is crucial for unravelling their fundamental biology. Exploring the genetic variation of the parasite can offer valuable insights into numerous aspects, and below are highlighted some specific areas where enhanced knowledge in this regard can provide novel perspectives.

2.4.1 Coinfection and Competition

Parasites rarely infect in isolation, and it is quite common to find hosts infected with multiple parasites, known as coinfections. Coinfections can involve parasites of different species or different strains/haplotypes within the same species.

When looking at pathogens overall, the result of coinfections generally depends on whether the parasite species/strains compete or cooperate, and how closely related they are. Competing strains tend to evolve greater virulence, since the faster a strain reproduces, the faster it can take advantage of limited resources (Nowak and May, 1994). However, closely related strains in a coinfection will tend to evolve to be less virulent, as the benefit from outcompeting each other wains due to kin selection (Frank, 1996) when individuals altruistically favour the reproductive success of close relatives (Griffin and West, 2002). However, when closely related strains cooperate in an infection, fitness and therefor virulence increases, as one would predict using kin selection theory (West and Buckling, 2003). This can be seen in siderophore (molecules that scavenge iron, a rare and limiting resource) producing bacteria (Harrison *et al.*, 2006) and RNA viruses that share RNA replication resources within a cell (Turner and Chao, 1999).

Avian malaria has provided rich studies on co-infections, as there are host species that are susceptible to multiple malaria species and strains of the same species, with well characterised disease progression. Palinauskas *et al.*(2018) demonstrated similar mortality when experimentally infecting canaries with *Plasmodium relictum* and *P. elongatum*, however because *P. relictum* targets mature RBCs and *P. elongatum* immature and mature RBCs, there was less

competition. In a study in Eurasian siskins, two different lineages of *P. relictum* were analysed. The more widespread lineage, SGS1, supressed the parasitemia of the specialised lineage, GRW4, though the overall mortality was not affected (Aželytė *et al.*, 2022).

Even within a strain, there might be competition for host resources between each haplotype (every parasite is haploid in the host), as it is a safe assumption that each new infection will begin with a population of parasites. This genetic variation provides an interface for selection, as each haplotype may adapt differently to the host which may results in competition between haplotypes. Since the parasites reproduce asexually within the host, the genetic material cannot be mixed through sexual reproduction to create new genetic variations. As a result, mutations are the only way for new genetic variation to arise in the parasite population within the vertebrate host.

However, the rate of mutation in parasites is generally low (Bopp *et al.*, 2013; Murray *et al.*, 2017), and the genetic variation that arises may not always be beneficial for the parasite's survival. Therefore, the selection pressure imposed by the host's immune system and other environmental factors is crucial in shaping the genetic diversity and evolution of parasite populations.

To study selection and competition within infections, crucial for host-parasite co-evolution, we require tools to examine genetic variation within species, alongside between species or lineages. These tools enable us to investigate the selection and competition of new mutations or haplotypes within individual hosts.

2.4.2 Immunity

Experiments using serum from *Plasmodium circumflexum*-infected birds demonstrated protection from the same parasite in uninfected canaries, and the more similar the strain, the better the protection (Manwell and Goldstein, 1938), which highlights the importance of understanding the genetic background of the parasites. One approach involves parasite versus host immunity. It is well documented that host susceptibility varies widely within and between parasite species (Bensch, Hellgren and Pérez-Tris, 2009), and one of the main host factors is the major histocompatibility complex (MHC) (Westerdahl, 2007), likely because the MHC genes are extremely variable in copy-number and form. They are part of the adaptive immune system which presents pathogen proteins to “surveillance” and B and cytotoxic T-cells. The

diversity of MHC genes means that different individuals within a host population have distinct MHC profiles, resulting in varying abilities to recognise and present antigens. This diversity poses a challenge for pathogens as they need to adapt and evade immune responses mounted by different MHC profiles. Pathogens that can escape recognition by one MHC profile may still be recognised and eliminated by individuals with different MHC profiles, driving ongoing co-evolution between pathogens and hosts (Westerdahl, 2007; O'Connor *et al.*, 2019). One example in blue tits showed that particular MHC-1 alleles provided partial resistance or susceptibility to Leucocytozoon infection depending on the combination of alleles (Rivero-de Aguilar *et al.*, 2016). Advancements in sequencing larger portions of the parasite genome hold the potential for future studies to not only identify parasite haplotypes/strains that can evade host recognition but also pinpoint the specific genes or proteins within the parasite that facilitate recognition by the host's immune system, including the MHC.

2.4.3 New Host Acquisition

Spill-over events (the acquisition and spread of a parasite in a new host species) are the major cause of novel disease outbreaks in animals and humans (Palmer, 2011; Smith, 2014). These zoonotic diseases tend to be more virulent than established pathogens, since there is no existing resistance or tolerance mechanisms in place (Jones et al, 2008). Avian malaria is notorious in this regard (Lowe *et al.*, 2000). It was responsible for multiple species of Honeycreepers in Hawaii after the introduction of the vector *Culex quinquefasciatus*i and the release exotic birds in the early 1900s (Riper *et al.*, 1986; Beadell *et al.*, 2006; Atkinson and LaPointe, 2009). In order for a parasite to jump to a new host however, there must be both genetic and ecological compatibility.

By definition, parasites are adapted to their hosts. However, in order to spread to new host species, they must have genetic resources to interface with the new species. The parasite must be able to evade/suppress the immune system and utilise the internal resources of the host at the molecular level. Generalist parasites presumably carry the genetic tools for this phenotypic plasticity which is a fitness advantage in the face of the extinction of their original host. If the plasticity is due to a large choice of genes for each host, there will be the cost of antagonistic pleiotropy (Futuyma and Moreno, 1988) because none of the environments are ideal. The alternative solution is alternative gene

expression, where the parasite optimises behaviour to each host species. This still requires additional machinery (host sensing, silencing, enhancing, alternative splicing, epigenetics), but if the parasite encounters new opportunities, it can take advantage of them, which is a strong selection for plasticity (Brown, Cornforth and Mideo, 2012). And this is especially true when parasites must exploit individuals within a genetically diverse single-species population (Kalbskopf *et al.*, 2021). Finally, parasite species with smaller genomes may still acquire and adapt to new hosts if the genetic resources are spread across (available as genetic variability within) the population. Natural selection will find the compatible parasite-host combination and then reproduction, horizontal gene transfer, or sexual reproduction will exploit the compatibility.

The genetic mechanisms underlying the adaptation of generalist and specialist parasites to new hosts remain largely unknown. It is crucial to investigate how generalist and specialist parasites differ in terms of genetic variability. Specifically, it is essential to determine whether generalist parasites generally exhibit greater genetic variation in their populations, potentially as a consequence or cause of their generalist nature compared to specialists. Additionally, exploring whether generalist and specialist parasites differ in the types of genes they possess is of significant importance. These questions highlight the need for more extensive research on the population structures and genomics of parasites to gain deeper insights into these genetic dynamics.

2.4.4 Transmission ecology

Avian malaria provides a useful non-human model for studying ecology and spread, because the socio-ecological variables that are unique to the human model become insignificant when dealing with parasites that affect wildlife (Garamszegi, 2011). There are multiple vectors with different niches and host preferences (Reed *et al.*, 2003). Migrating birds are exposed to different populations of parasites in their wintering and breeding sites, and their popular stopover sites provide enhanced opportunities for disease transmission within and between species (Garamszegi and Møller, 2007). Avian malaria parasites in the Europe-Africa flyways rarely transmit between breeding and wintering areas (Hellgren, Waldenström, *et al.*, 2007). There appears to be more phylogenetic diversity in the North-South American breeding and wintering grounds due to cross-site transmission (Ricklefs *et al.*, 2017).

Genetic compatibility is only the first requirement for host acquisition. The predominant theory on host specificity is evo-ecological in nature (Devictor *et al.*, 2010), (Poulin, 2011) which means that parasites must adapt to their hosts, which creates a tight link, an antagonistic partnership that evolves over time. Acquiring a new host species is a stochastic process and requires an intersection of physiology and interaction with the old host (Woolhouse and Gowtage-Sequeria, 2005), but recent research is showing environmental factors are more important shapers of host specificity than previously thought (Thrall et al., 2007, Nylin et al., 2018).

Here the concept of the ecological niche is adapted to host specificity. The potential resource availability of susceptible hosts can be restricted by the presence of a susceptible vector, for example, which are considered highly dependent on climate (Fecchio, Wells, Bell, V. v. Tkach, *et al.*, 2019). This is known as environmental filtering. When the parasite utilises potential resources, the niche is realised (Wells *et al.*, 2018). The contribution of environmental change in the form of shifting cycles (such as global warming) to the exposure of parasites to new hosts is a metatheory known as the Stockholm paradigm (Hoberg and Brooks, 2015).

There are two approaches for analysing host specificity. The first is trait-based, which uses factors like behaviour, environmental preference, body size, and phylogenetic relationships. Hosts that share common traits are more likely to share parasites (Webb et al., 2010). The second approach is the network of interactions between different hosts and parasites in an environment on the level of the community. Hosts that share many nodes in the network are more likely to share parasites (Poulin, 2010).

Before these theories can be thoroughly tested in natural settings, it is crucial to determine the species or subspecies boundaries of the parasites. Specifically, when identifying multiple hosts within a specific transmission area or niche, it is essential to ascertain whether we are dealing with a single species or multiple cryptic species, each with a narrower niche. As these parasites often exhibit cryptic morphology below a certain level of genetic diversity (Hellgren, Križanauskiene, *et al.*, 2007), more precise molecular methods are required to define the boundaries of "species" for these parasites. This, in turn, enables accurate assignment of ecological boundaries. Exploring the population structure of the parasites offers a viable approach to discovering and comprehending their genetic boundaries.

2.4.5 Population structure

The most obvious cause for a lack of genetic mixing between strains is geographical separation. Recombination of multilocus genotypes will not occur across physical barriers. Perhaps a less obvious but related hurdle for the association of different genotypes is demographics. If a vector where the parasite undergoes sexual reproduction can only carry a single genotype at a time, then there will not be opportunities for recombination between genotypes. The same is true for obligate intracellular parasites that will only have opportunities for recombination if there is coinfection of a cell. If the product of two genotypes has lower fitness, then hybrid incompatibility will prevent the presence of the intermediate genotype. Epidemics will reduce diversity because one genotype outcompeted all the rest.

There is a knowledge gap to be filled in understanding how genetic variation results in contrasting outcomes in virulence, host specificity, and rates of transmission, since these factors can vary widely between otherwise similar parasite species. One way to observe and identify different species or ecological entities in the wild is to study how different genotypes or haplotypes are sorted across geography or host species in the wild. For example, linkage disequilibrium (LD) is evidence that alleles are not randomly associating in a genome in an individual, but rather are found together more than one would expect by chance. Genome-wide patterns of LD lead to genetic structure in the population (Wright, 1969). It is easier to measure LD than identify its cause as there are multiple factors that effect it. And this is especially true for parasites, which have short generation times (relative to their host) and antagonistic relationships with their environment. However, because malaria is haploid in the host, where it is most likely to be sampled, LD and other classic statistical frameworks cannot be used. Instead, relatedness, measured through Identity by Descent (IBD) is substituted. Unlike Fst and nucleotide diversity (π) which rely on allele frequency, IBD uses the frequency of outcrossing recombination, which breaks down the chromosomal segments identity over generations (Taylor *et al.*, 2019), where whole genome data is generally needed (Neafsey, Taylor and MacInnis, 2021). Despite advancements in sequencing technology that offer improved genetic resolution of parasites, comprehending and studying their population structure remains challenging. One factor that complicates such investigations is the varying levels of clonal population structure observed in malaria parasites, attributed to selfing.

2.4.6 Clonality

Successful malaria transmission requires that the mosquito bites a host at least twice, first biting an infected host, then after sufficient incubation, a naïve host. A mosquito may bite multiple infected hosts in a row, providing the chance for interbreeding. However, depending on the population structure, outbreeding is not always possible, and more than two blood meals are not guaranteed. The parasite must form a zygote (male and female gametes fuse) to form the next infective form in the mosquito, so selfing may be a necessity if there is a lack of background genetic diversity. One of the central aims of this thesis is to elucidate the degree of genetic diversity within an infection.

Clonality can occur due to mitotic division or mating through selfing and homogamy (Tibayrenc and Ayala, 2002). However, clonality should not be confused with monomorphism (few if any polymorphisms compared to the reference). To identify a population as clonal, it must have a distinct genotype that is genetically distant from related populations and remains steady over time and space. Historically, virulent viral, bacterial, fungal and protist strains were defined as species because they have these clonal population structures (Tibayrenc and Ayala, 2012). Testing for clonality requires two successful tests. The first test is a strong LD signal from many high resolution markers (Awadalla, 2003), commonly measured with the Index of Association (Smith *et al.*, 1993) or perhaps the *g* test which tests for LD correlation between otherwise independent loci. Secondly, when analysing 'core' genes, genes without strong selective pressure, the clonal population should segregate phylogenetically (Tibayrenc and Ayala, 2012). Clonet has been suggested as the term for a population that is defined by the genotype of a single marker (Tibayrenc and Ayala, 2002).

The sampling method will determine the type of conclusions one can draw. Long term and wide-spread sampling reveal genetic variation stability (or lack thereof) for a population (Tibayrenc, 1995). It may fall to Wahlund bias[i], however (de Meeûs, Lehmann and Balloux, 2006). Deep, small-scale sampling reveals the state of the population in an instance of time and is more resistant to the Wahlund effect.

Clonality can be seen as the stabiliser and exploiter of sex in many successful pathogen superspreaders. Clonal parasites can carry beneficial multilocus genes without the burden of recombination or outbreeding (Michod, Bernstein and Nedelcu, 2008) which increases the effects of natural selection (Avise, 1994). However, parasites appear to carry genetic machinery that maintains a

balance between clonality and recombination, the latter of which creates the possibility of new genic combinations (Tibayrenc and Ayala, 2012).

2.4.7 Population effects and Virulence

Studying the population structure of the parasite is crucial for comprehending the dynamics of an outbreak in a new population, which includes factors such as host population size, connectivity, susceptibility levels, and epidemiology. These insights significantly impact the trade-off between virulence and transmission within that particular system. When a pathogen is introduced into a susceptible population, the abundance of hosts favours rapid spread with high host mortality, so the parasite can have its cake and eat it too. But once the population is more resistant, the pathogen must be more subtle to spread (Berngruber *et al.*, 2013). If the pathogen is introduced to smaller clusters within a population with lower host movement between clusters, it can spread well within the cluster, but then find its unable to spread if it kills the cluster. This is known as self-shading (Boots and Sasaki, 1999), which would lead to a stronger virulence/transmission trade-off and can be seen in phage experimental systems (Berngruber, Lion and Gandon, 2015).

There are many studies (Razakandrainibe *et al.*, 2005; Kiwuwa *et al.*, 2013; Duffy *et al.*, 2017, 2018; Messerli *et al.*, 2017; Wahlgren, Goel and Akhouri, 2017; Tessema *et al.*, 2019; Osborne *et al.*, 2021; Coulibaly *et al.*, 2022) analysing the population structure of human malaria especially directed at understanding the spread of drug resistance and the emergence of epidemic strains in new habitats. The population structure of *P. falciparum* was revealed on a global scale using deep sequencing by Manske *et al.* (2012) where they identified many more unique SNPs and within-host diversity in regions with higher endemicity compared areas with lower rates of transmission.

Recently, a method called Selective Whole Genome Sequencing (SWGS) has been utilised to examine the populations of *P. falciparum* (Oyola *et al.*, 2016) and *P. vivax* (Cowell *et al.*, 2017). This approach allows for the analysis of a broader range of genes and markers from natural infections, rather than solely focusing on specific PCR-targeted genes.

Avian malaria remains a topic of extensive population genetics research, yet our knowledge about it is limited (Humphries, Stacy and Ricklefs, 2019; Videvall, 2019). To advance the field and utilise it as a model system for studying population genetics and the evolution of virulence, there is a crucial

requirement for enhanced tools and theoretical frameworks for conducting genomic studies on a broader scale. Achieving this milestone would unlock a wealth of opportunities for future investigations, given the system's diversity in terms of host specificity, transmission areas, and variations in virulence.

[1] Wahlund bias, also known as the "substructure bias" or "stratification bias," is a statistical phenomenon that arises in population genetics and epidemiology studies, particularly in the context of genetic association studies. It occurs when the study population is not homogeneous but is instead divided into subpopulations or strata that have different allele frequencies for the genetic variants being studied.

3 Aims and objectives

The aim of this thesis was to create methods allowing me to characterise the intraspecific genetic variation of the SGS1 lineage of *Plasmodium relictum,* avian malaria, and study how this variation might affect infection outcome. We approached this with the following objectives and questions.

1. **Understand the variability in expression of the parasite between hosts over time.** We attempted to characterize and understand how the expression of the parasite changed in relation to parasitemia over time, as there is a distinct rise in parasitemia at around 8 days post infection and a strong decrease from 16 days post infection. We sequenced and compared the mRNA of the parasite at points before and after peak parasitemia. The aim was to identify the function of differentially expressed genes and genes associated with extremely high variation in expression using GO (Gene Ontology) terms and metabolic pathways as the parasites adapt to genetically diverse hosts.
2. **How much genetic variation is there in the European SGS1 population?** The MSP-1 nuclear gene evolves faster than mitochondrial genes, potentially revealing genetic variation below the level of mitochondrial lineages. The MSP1 gene was sequenced from infections in birds across Europe to establish the population structure of SGS1 within this geographical region.
3. **Develop and verify a method to identify the nuclear genetic variation of *Plasmodium relictum* (SGS1, GRW4, GRW11 lineages).** Because genomic scale sequencing of avian malaria is challenging, we used targeted exome sequencing to enrich for parasite material. The aim was to develop and verify this technique in *P. relictum* and describe some of the variation in this species using 25 nuclear genes.
4. SGS1 samples from a crossbill and sparrow were inoculated separately into two cohorts of canaries and resulted in low and high mortality respectively. **Why does infection outcome vary when the species of the host donor of infection is different? What is the inherent**

genetic variation in the two infections? It was speculated that the two strains of SGS1 were genetically different, and sequenced the mRNA from the infections in the canaries to identify the genetic variation. Then we tried to identify the genes behind the differences in infection outcome.

5. Our previous experiments show much more genetic variation within the lineage of SGS1 than seen before. **How much diversity is there in a single infection? How does that diversity change over time within an infection? Can the diversity be linked to infection outcomes?** In the same differential expression experiment as used in Objective 1, I sequenced the targeted regions of the parasite genome, using the method developed under Objective 3, at the same time points, and characterised the majority haplotypes for each sample.

4 Methods

4.1 Study Species

4.1.1 *Plasmodium relictum*

Haemosporidian parasites are globally dispersed protists that require a dipteran vector to spread. The avian parasites can be found in three different genera: *Leucocytozoon* (transmitted by black flies), *Haemoproteus* (transmitted by biting midges and flies), and *Plasmodium* (transmitted by mosquitos). The *Plasmodium* group infects 1877 unique species of birds (Bensch, Hellgren and Pérez-Tris, 2009) as of May 2023. The morphologically defined species of *Plasmodium relictum* is the most geographically widespread of these and infects the greatest variety of hosts (Bensch, Hellgren and Pérez-Tris, 2009) and is on the list of the 100 worst invasive species (Lowe *et al.*, 2000). Thanks to the widespread use of the mitochondrial *cytochrome-b* gene marker, which can be easily barcoded using PCR on blood samples, mitochondrial lineages can be easily identified. The most common lineage of *P. relictum* is SGS1 (sometimes represented as pSGS1, the p indicating the family *Plasmodium*) and is the main focus of this thesis. *P. relictum* also includes GRW4, GRW11, LZFUS01, and PHCOL01, with only the first 2 included in this thesis, in Paper 3. GRW4 has gained notoriety when it caused the extinction of several native honeycreeper species after its introduction to Hawaii (Riper *et al.*, 1986; Atkinson and LaPointe, 2009).

There are extensive studies on *P. relictum's* "cousins" that infect humans (*P. vivax, P. ovale, P. malariae, P. knowlesi* and especially *P. falciparum*) and rodents (*P. vickei, P. yoelii* and the main model organism for mouse malaria: *P. berghei*). But they come with particular restrictions. Humans cannot be used for infection experiments, and the rodents used tend to be highly inbred, and therefore genetically 'simple' study systems. This may be by design, as one can use specific mutant mice lines that lack specific immune functions or host receptors. But if one wants to understand the evolution of malaria 'in the wild', where it is under selection of climate change, shifting host genotypes, changing vector abundances, while still being able to do controlled infection experiments, then avian malaria is an excellent option.

4.1.2 SGS1

SGS1 is a mitochondrial lineage of *P. relictum* that has been found in 146 bird species, comprising 12 orders on every continent except Antarctica (which lacks its vectors) and Oceania. SGS1 has further been detected in 9 species of mosquito (Bensch, Hellgren and Pérez-Tris, 2009). Its virulence varies extensively across different host species (Meister *et al.*, 2021) and can even differ within the same species (Kalbskopf *et al.*, 2021, manuscript., no date) which has made it subject to extensive infection experiments, especially since it infects both model and non-model organism bird species that are easy to maintain in aviaries. The infection experiments have been used to understand disease progression (Palinauskas *et al.*, 2008, 2020), immune response of the host (Videvall *et al.*, 2020), and interactions of SGS1 with other *Plasmodium* infections using coinfection methods (Palinauskas *et al.*, 2018; Aželytė *et al.*, 2022).

Table 1: Nucleotide differences in the cyt. b gene for the 4 lineages in this thesis. PCR covers 480bp, the whole gene is 1130bp.

	GRW11	GRW04	SGS1	DONANA05
GRW11		10	1	2
GRW04	10		11	12
SGS1	1	11		1
DONANA05	2	12	1	

Though *P. relictum* in general has a very wide distribution, SGS1 was long thought of as an Old World parasite (Chagas, Harl and Valkiūnas, 2021), however, recently it was detected in non-migratory North (Theodosopoulos *et al.*, 2021) and South American birds (Marzal *et al.*, 2015) which is one of the reasons it's considered to be extremely invasive. GRW4 seems to be predominantly transmitted in tropical regions in the Old World, whereas it can be found to be transmitted in temperate regions in the New World and exhibits reproductive isolation between the other lineages of *P.* relictum when examining the *msp1* gene (Hellgren *et al.*, 2015). In terms of nucleotide differences, GRW4 displays 11 nucleotide differences while GRW11 has only 1 compared to SGS1 (Table 1) at the mitochondrial gene *cyt b*. These findings further strengthen our perception that GRW4 represents a distinct species. In

this thesis, we contribute to the existing research by analysing a broader range of genes.

P. relictum is one of two published avian malaria genomes, the other being *P. gallinacium* which was published at the same time (Böhme *et al.*, 2018). This has opened genomic studies of avian malaria, providing valuable insights into the genetic makeup, evolution and greatly simplified the bioinformatics required to study it. As a result of a misidentification during the initial infection assessment, the sequenced genome was initially believed to belong to SGS1. However, a single nucleotide variation in the *cyt. b* gene indicated that the sample was actually DONANA05. Therefore, the genome is now designated as SGS1-like due to its similarity to SGS1 with this minor difference.

Böhme et al. accomplished the generation of sufficient sequencing material for the parasite genome sequences by dissecting mosquito infected midguts. The genome sequence was employed in constructing phylogenomic trees to analyse the evolutionary relationships within the *Plasmodium* genus. They discovered that the bird parasites occupied a distinct position outside the *Plasmodium* clade that infects mammals. Moreover, certain genes previously assumed to be exclusive to the primate-infecting *Laverania* clade were identified within the genomes of these avian parasites. In terms of avian *Plasmodium* genomes, they contained tens of genes and multiple transposable elements that are unique to them. These findings present potential future targets for investigating the molecular mechanisms underlying mammal-specificity and bird-specificity within *Plasmodium* (Videvall, 2019).

4.2 Direct Inoculation

With the exception of the samples obtained from natural infections in Paper 2 and 3, all of the sequenced material originated from birds that were infected with blood sourced from previously infected birds. This process is called direct inoculation and omits the use of the vector.

Blood is taken from the donor host during high parasitemia and injected into the muscle or vein of the recipient. This has several advantages over using the vector: Blood can be stored frozen and thawed when it's needed; mosquitos are more complicated to maintain and use; one can quantify the number of meronts and ensure all the recipients receive the same dose; not all the parasites

from a mosquito bite will enter the blood stream, but may get stuck in the dermis or drain away in the lymph; and depending on the parasite species, the vector is not always known, so in some cases direct inoculation is the only option for experimental infections. However, this method deviates from natural vector mediated infections which means certain caveats must be considered. Mosquitos infect their hosts with sporozoites, not meronts, so the primary exo-erythrocytic cycle, which will change the immune response. Vectors also impose their own selective pressures on the parasite reproduction and transmission, which will change the parasite haplotypes that are transmitted to the host compared to direct inoculation. (Palinauskas *et al.*, 2020)

When studying the host's response to the parasite (not directly studied in this thesis), controls must be used where uninfected blood is injected into the host, as foreign blood will stimulate a response from the host.

4.3 Sequencing Methods

4.3.1 Challenges when studying Endocellular Parasites

Avian malaria parasites are rarely found outside a cell when in the host. This makes capturing parasite genetic material without high levels of host contamination almost impossible, since the host genome is approximately 50x bigger than the parasite's genome (Host=1.2Gbp, *P. relictum*=22Mbp). And unlike mammalian red blood corpuscles (RBCs), avian erythrocytes are nucleated. There are methods being developed that attempt to circumvent this issue.

Cell sorting techniques may enrich for infected blood cells, but they may bias the collection of stage-specific parasites (unpublished data), which is important when doing expression analysis. The current *P. relictum* genome was sequenced by extracting oocysts from the dissected guts of infected *Culex* mosquitos (Böhme *et al.*, 2018). This method works well for establishing a reference genome, but does not scale or isn't applicable for use in large ecological or experimental analyses.

This project uses two more methods. We sequenced the mRNA of infected host blood, with the presumption that the host-parasite ratio is more in the parasites favour. It was also sequenced at great depth to ensure enough

coverage. The other method is targeted sequencing of the parasite genome using exome capture probes that were also designed to avoid binding to the host genome. Both of these methods are reliant on high levels of parasitemia, therefore experimental infections of susceptible hosts were used to develop the experimental and bioinformatic methods.

4.3.2 RNA-seq

mRNA sequencing is a powerful technique used to analyse the transcriptome of a parasite. The process begins with the isolation and purification of RNA from biological samples. The isolated RNA is then fragmented and used as a template for cDNA synthesis, where reverse transcription generates complementary DNA molecules. These cDNA fragments are further processed to create a sequencing library by adding adapters. The library is then sequenced using the high-throughput Illumina HiSeq platform, generating millions of short reads. The generated reads are aligned to a reference genome to determine gene expression levels and identify expressed genes. Data analysis, including differential expression analysis and functional annotation, is performed to interpret the results and gain insights into biological processes.

RNA-seq was vital in the early exploration of the genic content of avian malaria. In 2013, Hellgren et al were the first to sequence the RNA from the blood of an avian malaria infected host and retrieved the sequence for the cell invasion MSP1 gene, which was then used for PCR assays. The cell invasion and immune related genes in *P. gallinaceum* (fowl malaria) were studied using transcriptomes in experimentally infected chickens in 2014 and 2015 by Lauron et al who identified novel SNPs and large differences in gene family expansions when compared to *P. falciparum*. The first nearly complete avian malaria transcriptome used *Plasmodium ashfordi* to preform differential expression analysis and revealed plastic responses of the parasite to their hosts (Videvall *et al.*, 2017).

4.3.3 Targeted Sequence Capture

Targeted sequence capture is a technique used to selectively sequence and analyse specific regions of the genome that encode proteins, known as the exome. The exome represents only a small percent of the entire genome but is responsible for all the protein in the cell.

Targeted exome sequencing involves two main steps: target enrichment and sequencing.

Target Enrichment: In this step, specific regions of interest in the genome are captured or enriched for sequencing. We specifically chose genes that have orthologous copies across 17 different apicomplexan parasite species. Additionally, we selected genes associated with cell invasion biology, mitochondrial genes, and a random selection of genes. The technique we used is called hybridization capture. It involves the use of custom-designed probes or baits that are complementary to the exonic regions of interest. These probes selectively bind to and capture the desired genomic regions. The probes were designed to not overlap each other.

Sequencing: After the target enrichment, the captured DNA fragments are sequenced using the high-throughput sequencing platform Illumina MiSeq. The captured DNA fragments are fragmented further, if necessary, and ligated with sequencing adapters to generate sequencing libraries. These libraries are then loaded onto the sequencing instrument, where both strands undergo amplification and sequencing-by-synthesis reactions to generate millions of short sequence reads.

4.4 Bioinformatics

4.4.1 Trimming, QC, Filtering host reads

A summary of the bioinformatic methods used here can be seen in Figure 4-2.

Illumina short read sequencing requires various quality control processes to ensure accurate and consistent output. These are the basic steps we followed for both RNA-seq and targeted exome sequencing, with differences pointed out.

Raw Data Assessment: The first step is to assess the raw sequencing data files (in FASTQ format) to identify potential issues. This involves examining basic metrics such as the number of reads, read length, and base quality scores. It helps identify any anomalies or irregularities that may affect the subsequent analysis steps. FastQC (Andrews, 2010) produces statistics for each sample and then MultiQC (Ewels *et al.*, 2016) collates each sample into a single file for easy comparison.

Adapter Trimming: Illumina sequencing involves the use of adapters, which are short DNA sequences attached to the ends of the fragments being sequenced. These adapters need to be removed before further analysis.

Quality Filtering: Quality filtering involves removing low-quality reads based on their base quality scores. Each base in a read is assigned a quality score, which represents the confidence in the accuracy of that base call. By applying a quality threshold, reads with low-quality bases (typically defined by a specific score or Phred score threshold) are filtered out, reducing the likelihood of including erroneous data in downstream analysis. Both the adapter and quality filtering was done by Trimmomatic (Bolger, Lohse and Usadel, 2014). The RNA-seq data was also filtered for quality by the Illumina chastity filter before we received it.

Contaminant Removal: Bird DNA or RNA are a significant source of contamination. For Papers 4 we processed the fastq files using FastqPuri (Pérez-Rubio, Lottaz and Engelmann, 2019). The filter in FastqPuri uses the *Serinus canaria* genome to create a 'database' which it compares to the reads in the fastq files. Any reads that are too similar to *Serinus canaria* genome are removed from further analysis.

4.4.2 Mapping, Differential expression, Consensus sequences

The reads that pass the QC filters are mapped to the reference genome. Different versions of the *P. relictum* SGS1-like genome were released over the course of this PhD. The main discrepancy occurred when sequencing results from probes designed for an early version of the genome were mapped to version 54, resulting in differences in gene annotations.

NextGenMap (Sedlazeck, Rescheneder and Von Haeseler, 2013) was selected to map the genomic DNA reads because it is optimised for mapping to polymorphic genomes, which is particularly important in Paper 3 which involves mapping reads from the more distant *P. relictum* GRW4 lineage.

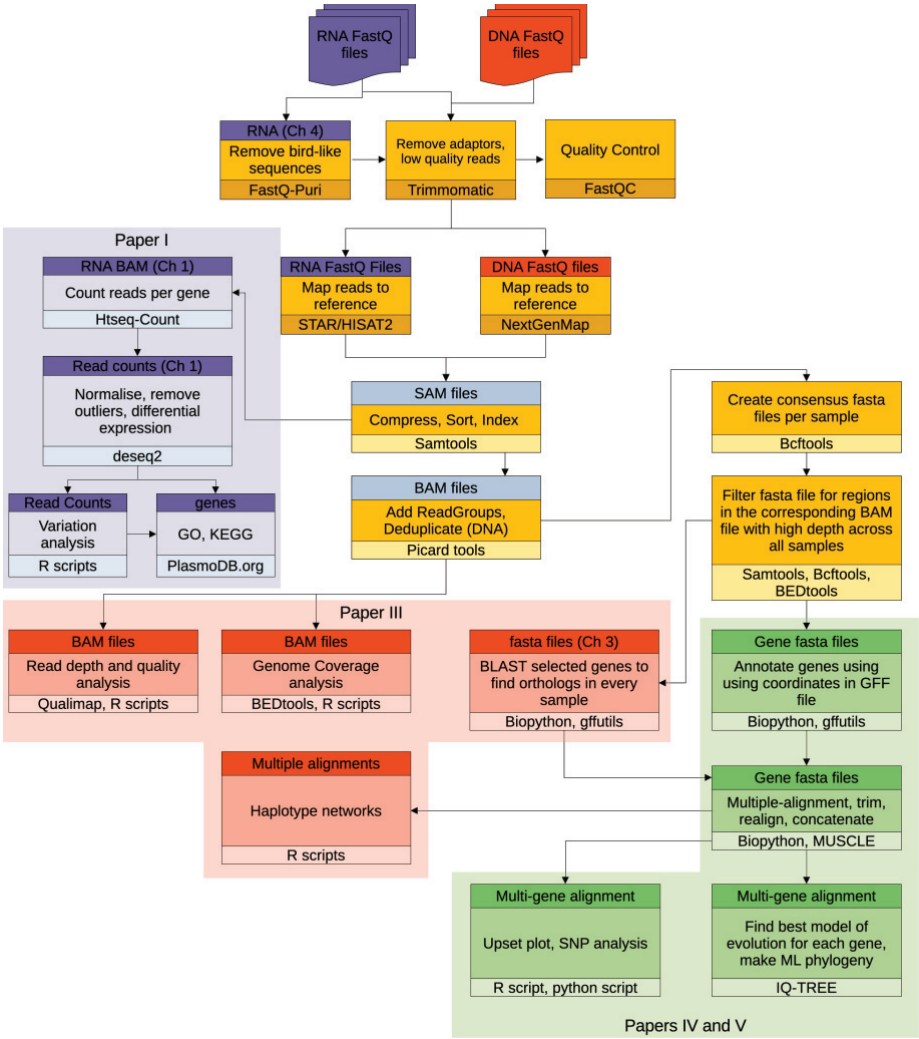

Figure 4-1: A broad overview of the bioinformatics pipelines used in this thesis. Paper I is in the purple box, Paper III in the red box, Papers IV and V in the green box. Paper II used some samples from Paper III and custom methods described elsewhere. The darker box is the file type, the middle lighter box is the process, and the lighter bottom box is the software used.

RNA-seq mapping requires a splice aware mapper, which takes introns into account when mapping reads. STAR (Dobin *et al.*, 2013) was used for the

differential expression experiment (Paper 1) and HISAT2 (Kim *et al.*, 2019) was used for Paper 4.

During the sequencing process, PCR duplicates can be generated, leading to artificially inflated read counts. Duplicate removal identifies and removes duplicate reads after mapping to prevent bias in downstream analyses, ensuring accurate representation of the original sample. We used MarkDuplicates from Picard tools (*Picard Tools - By Broad Institute*, no date). This step was not performed for the RNA-seq data, as one expects multiple identical copies for mRNA data, and removing copies would depress the counts for highly expressed genes.

To preform differential expression analysis in Paper 1, the number of reads for each gene must be counted. HTSeq-count (Anders, Pyl and Huber, 2015) does this by using the coordinates in the BAM file and the corresponding gene in the gene annotation file (GFF) for each sample. Deseq2 (Love, Huber and Anders, 2014) assembles the gene counts into a matrix and transforms the counts across all the samples. Then extreme outliers are automatically removed, and the testing is performed to find genes that expressed differently in the two groups.

Paper 2 focused on one gene, *msp1*, and used samples from wild-caught infections and the same samples used in Paper 3. Because parasitemia in wild-caught infections tend to be very low, PCR amplification of a smaller block of *msp1* was necessary for these samples. They were sequenced using Sanger sequencing. The samples from Paper 3 (experimental infections) were targeted with capture probes and sequenced using Illumina MiSeq. For each sample, a consensus sequence was generated for the gene of interest using Genious Prime and then aligned using MUSCLE (Edgar, 2004).

Papers 3, 4 and 5 relied on making consensus genome sequences for each sample. BCFtools (Danecek *et al.*, 2021) used the reference genome and called SNPs in the BAM file to create the majority consensus sequence fasta file.

We only wanted regions of the genome that were supported by enough depth to make high confidence SNP calls, so SAMtools (Li *et al.*, 2009), bedtools (Quinlan and Hall, 2010) used the depth information in each BAM file and created one bed file that contained common regions in every sample, which was used to filter each sample's reference genome. This resulted in one fasta file per sample with coordinate- labelled headers for each region.

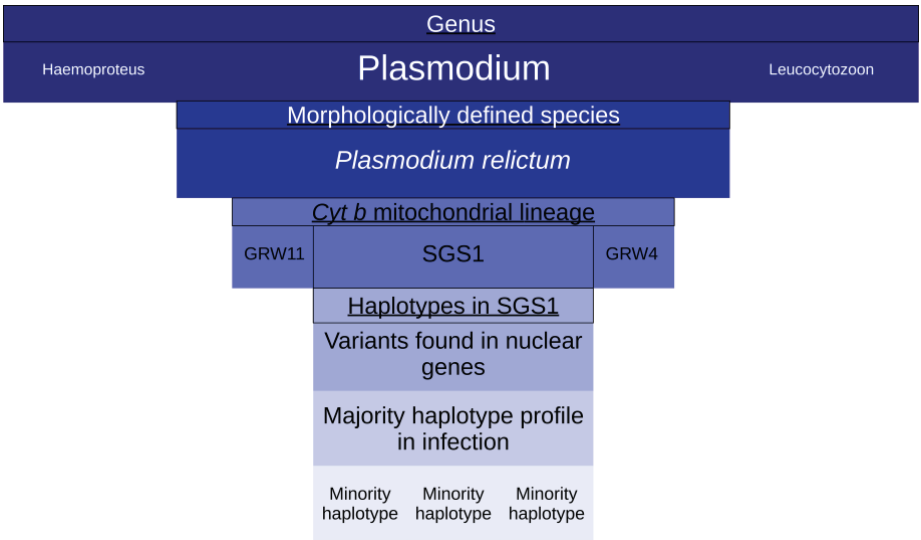

Figure 4-2: The *Plasmodium* hierarchy, increasing in genetic resolution as it descends. Papers 1, 3, 4, and 5 examine the majority haplotype profiles in each infection, and Paper 2 and 3 compares nuclear genes within the SGS1 lineage and between lineages. Currently, the minority haplotypes and fully phased haplotypes are not under consideration due to the limits of the methods used. See Conclusion (History and Caveats).

Because we're using short reads, we cannot link the SNPs to individual parasites, so the process of creating the genome from the BAM file for each sample is in effect selecting a mix of most abundant haplotypes in that sample. Therefor each sample represent a profile of the population of parasites in that sample (Figure 4-2).

4.4.3 Multiple alignments, Phylogenies, SNPs

The methods used to attain phylogenies and SNPs were improved and refined from Papers 3, 4 to Paper 5. As Paper 3 was primarily a method paper, we selected 25 nuclear genes that were used by the Huang *et al* (2019) analysis to verify the method. Once the consensus sequences were attained, we used BLAST to find the genes in the genome. Papers 4 and 5 used the genomic coordinates and gene annotations to select all genes with enough coverage which resulted in hundreds more genes.

The alignments were created over multiple steps: 1) align each gene, 2) trim start and end of gene for gaps, 3) realign, 4) quality control. MUSCLE was selected to perform the alignments as it was the fastest and most consistent after checking the alignments manually.

Each alignment was concatenated using Biopython into a partitioned nexus file which IQ-TREE (Nguyen *et al.*, 2015) can read. IQ-TREE used ModelFinder (Kalyaanamoorthy *et al.*, 2017) to select the best scoring model of evolution for each partition and then generated the phylogeny based on the maximum likelihood with 1000 ultrafast bootstraps.

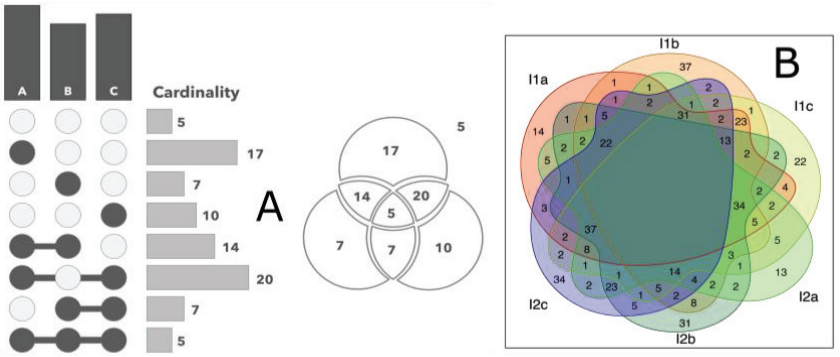

Figure 4-3: UpSet vs Venn. **A:** A simplified example with only sets of elements resulting in 8 intersections, shown as both an UpSet and Venn diagram. **B:** A Venn diagram of the data for the UpSet plot in Paper 4 which has 6 sets with 57 possible intersections.

A concatenated fastal file was created for the SNP analysis in R. We wanted to identify the SNPs that made each group of samples unique, so after removing all the gaps an UpSet plot (Conway, Lex and Gehlenborg, 2017) was created. This plot revealed unique intersections of SNPs present in each sample. The sample can be thought of as a set of SNPs.

We chose the UpSet plot because it is better than a Venn diagram to visualize intersections when there are more than 3 sets. Venn diagrams become unwieldy an increase in individuals, as exemplified in: Figure 4-3B, which is the same data used in Paper 4.

The UpSet plots revealed large intersections of SNPs that distinguished inoculation groups (Paper 4) and samples that deviated from the general parasitemia profiles (Paper 5). This prompted the exploration of the genes associated with these SNPs and an analysis of their respective functions using GO terms.

4.5 Analysis Summary

4.5.1 **Paper 1**

We infected 9 siskins with SGS1 and monitored their parasitemia every 4 days and took blood samples for RNA extraction on day 8 and 20 post infection (dpi). We were concerned with two primary outcomes in this experiment: 1) how does expression change over time between peak parasitemia at 8 dpi and low parasitemia at 20 dpi and 2) can we see differences in expression at the same time point related to high or low parasitemia. This is a straightforward analysis, which DESeq2 is designed for. In the first case, 2 birds (2 samples from each time point) were removed because the read depth was too low, but this still resulted in significantly differentially expressed genes.

In the second case, we tested for differentially expressed genes for the samples in day 8 that had high or low parasitemia. The day 20 samples had very low variation in parasitemia between the infected individuals and were therefor not used in the analysis. The parasitemia DE (differential expression) analysis resulted in only one significantly differently expressed gene, however the Principal Component Analysis (PCA) revealed a significant correlation between parasitemia and expressed genes. We found that parasitemia was correlated with the 2nd component of the PCA allowing us to use the genes most influential on the 2nd component for GO terms analysis.

When conducting the initial analysis, we observed large change in the variation in expression between the different time points. The PCA of normalised counts shows that the variation in 8 dpi samples is much smaller than in the 20 dpi samples, as measured by the area of the clusters for each group. We surmised that the genes that respond to the host will also be changing in expression more. So for each time point, we calculated the coefficient of variation (CoV) and plotted it in a violin graph, which confirmed what we saw in the PCA. Then the genes with the largest difference in CoV for each time point were selected

for GO term analysis, as they are assumed to be changing the most as a response to their individual hosts.

A significant challenge was the low sequencing coverage in the samples with low parasitemia at 20 dpi, which meant some of those samples had to be removed from the analysis, which reduced the statistical power. The large differences in sequencing depth between samples also creates issues when comparing expression between time points. As a solution, DEseq2 uses un-normalized gene counts but applies an rlog transformation to the data. Genes with very low counts across all samples were also removed from analysis as a form of independent filtering by DEseq2, which reduced the number of tests and the effect of multiple testing correction.

4.5.2 Paper 2

Prior to sequencing the genome of *P. relictum*, enough mRNA material was sequenced to obtain the sequence for the merozoite surface protein 1 (*msp1*) gene which meant PCR primers could be used to screen samples collected in the wild. The *msp1* gene is involved in cell invasion and immune evasion, making it polymorphic which is useful for characterising the population structure (Mwingira *et al.*, 2011; Hellgren *et al.*, 2015; Somé *et al.*, 2018). In this paper we aimed to characterise the population variation of SGS1, GRW11 and GRW4 using *msp1* in the palearctic region.

The sequences for 75 samples (36 host species) from the palearctic region were gathered from the Sanger sequencing of the PCR amplification and targeted exome sequencing from samples in Paper 3. The PCR probes targeted block 14 (269bp) of the *msp1* gene (4740bp) since the low parasitemia present in natural infections precludes sequencing the entire gene using this method. However, the targeted exome sequencing was successful in capturing the entire gene in some samples allowing for validation that no more variation was observed outside block 14. Double infections are common in avian malaria, so the *cyt.b* gene was included in all the analysis to filter out these samples which would otherwise complicate the *msp1* haplotypes. The sequences were then aligned using MUSCLE and the SNPs marked as synonymous or non-synonymous. The multiple alignment was used to create a phylogenetic tree using Geneious and the haplotype network was constructed using popart (Leigh and Bryant, 2015).

4.5.3 Paper 3

The aim of Paper 3 was to design and validate the method of a targeted exome capture protocol using probes designed for specific genes in the *P. relictum* genome. The basic technique is covered in the Targeted Exome Capture section above, and the details are explained in the paper itself. An important analysis, however, was understanding how parasitemia and genomic divergence affects coverage (breadth of genome sequenced) and depth (number of reads mapping to any particular region). We used 2 different sources of SGS1 experimentally infected into multiple hosts and one GRW4 (more divergent) and one GRW11 (less divergent) experimental infections, all with wide ranging parasitemia levels.

To analyse the coverage and depth we used multiple methods. Samtools stats, Qualimap and bedtools produced data that represents depth of coverage over targeted nucleotides, measured as the number of times a nucleotide was covered by a read. The breadth and depth were visualised for each chromosome over its entire length. We also produced rarefaction curves which help guide how deep one should sequence samples in future. Rarefaction curves are commonly employed to assess the extent of diversity within a population through the examination of the frequency of new individuals encountered during subsampling at different levels. If additional subsampling reveals the presence of previously unseen individuals, it indicates that further sampling is required. Conversely, if no new individuals are discovered, it suggests that the sampling saturated the population, capturing the entirety of its diversity.

For each sample, we called SNPs and extracted a consensus genome sequence. 25 genes were chosen from a previous *Haemoproteus* sequence capture study, and were BLASTed against each sample. The best BLAST hit for each gene was used to create multiple alignments which were manually inspected and trimmed. Then we constructed haplotype networks for 4 genes and all 25 concatenated genes. Two of the four genes were selected because they are membrane bound and interact with the host, and the other two are mitochondrial genes and provide insight into mitonuclear discordance.

4.5.4 Paper 4

Two groups of birds were infected with SGS1 obtained from two different sources, the first with blood from an infected cross bill, and the second a house

sparrow. The two groups (designated I1 and I2 respectively, for inoculum 1 and 2) showed differences in parasitemia, mortality, haematocrit levels and other health parameters, which were analysed every 4 days. These health parameters were compared between the groups and compared to their respective controls.

This paper describes the genetic variation present within SGS1 infections, both within and between the two groups, using RNA-seq and the same consensus genome technique used in Paper 3. However, compared to Paper 3 we extracted many more genes, relying on the annotations and coordinates to find gene regions across all the samples. To ensure we selected regions that were well supported by sequencing reads, we used depth information for each sample and only selected the regions of consensus genome that had at least 15 reads of depth across all samples.

Then each gene was aligned, trimmed, realigned, underwent quality control, and concatenated into a partitioned multiple alignment file. IQ-TREE constructed the maximum likelihood phylogenetic tree, treating each partition separately. Then the genetic variation in the alignment was analysed at the level of SNPs. We analysed how the SNPs were distributed between samples using an UpSet plot, and further investigated the functions of the genes the group-distinguishing SNPs were found in.

4.5.5 Paper 5

This paper refined the techniques used in Paper 3 and 4 but applied them to a larger sample set and considered time as a factor to see how the population of haplotypes within infections changed over the course of an infection. We used the exome capture protocol on the samples from Paper 1, sequencing part of the genome from 8 and 20 days post infection, as well as a few samples that did not meet the standards for RNA-seq in the infection experiment. This allowed us to see whether the composition of parasite haplotypes where different in each bird, although infected from the same donor infection, and whether the haplotype population changed between time points within infections.

Not all the samples could be sequenced using the exome capture protocol due to low parasitemia, so we couldn't achieve a complete one-to-one mapping of samples between papers.

We created consensus genome sequences for each sample, aligned all the regions with sufficient coverage across all the samples, and used those multiple alignments to create the phylogenetic tree and UpSet plot. The process of selecting regions across all the samples reduced the coverage used in the analysis from the targeted 2.5 Mbp to 1.03 Mbp, and after filtering the multiple alignments to remove alignments without SNPs and low-quality alignments, we were left with 53kbs across 120 genes that were comparable between samples.

5 Results and Discussion

5.1 Paper 1

5.1.1 **Parasitemia**

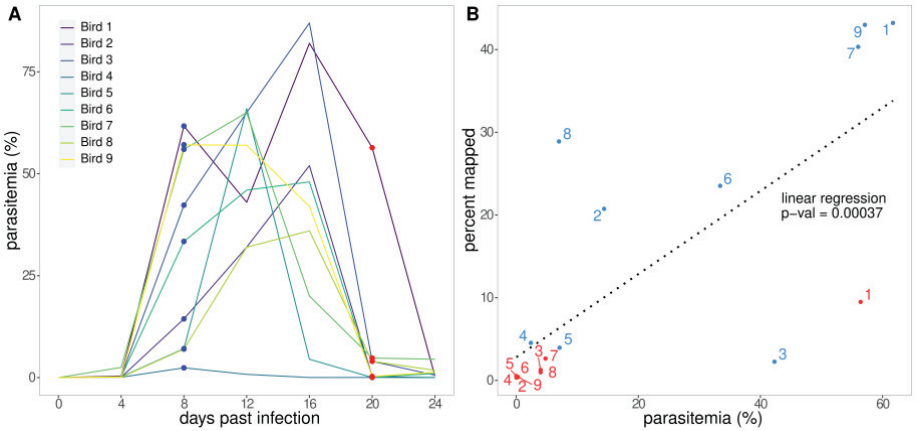

Figure 5-1: Parasitemia over time and its relationship to mapping success. **A**: Parasitemia over the course of the infection, typically reaching peak levels around 12 and 16 dpi. The blue and red dots represent when the birds were sampled for RNA extraction. **B**: Parasitemia is positively correlated with the number of reads mapped though there is wide variance.

As is typical for SGS1 infections, parasitemia peaks between 8 and 16 dpi, and falling parasitemia results in much lower sequencing and genome mapping success (Figure 5-1). The gene expression pattern closely follows time points, which can be seen in the cluster dendrogram (Figure 5-7) and the PCA of the gene expression counts. The exceptions are D8B3 (sample from day 8, bird 3) and D20B1, which buck this trend. In a similar experiment with *P. ashfordi* in siskins, parasite expression patterns were more stable over time but showed greater changes between individual hosts (Videvall *et al.*, 2017).

5.1.2 Differential Expression

The differential expression analysis between time points revealed 21 significant genes, 12 genes upregulated in the D20 samples, and 8 genes upregulated in the D8 samples. The upregulated genes in the D8 samples were related to cell movement and cell invasion (reticulocyte binding protein, fam-e protein family), which makes sense in the context of the early infection period, where the immune system has not yet effectively suppressed the infection, and the parasites can rapidly infect an abundance of red blood cells. In the latter time point (D20), parasite reproduction is lower, likely due to an improved immune response of the host. An issue we faced is the lack of annotations for a third to half of the significant genes, so there is still more to extract from this data set when more biological information is available. The genes upregulated in the D20 samples were generally related to cell metabolism and cell signalling, which may be a side-effect of the relatively low expression of these genes in the D8 samples compared to cell invasion genes, as cell metabolism may be down regulated during host-cell invasion.

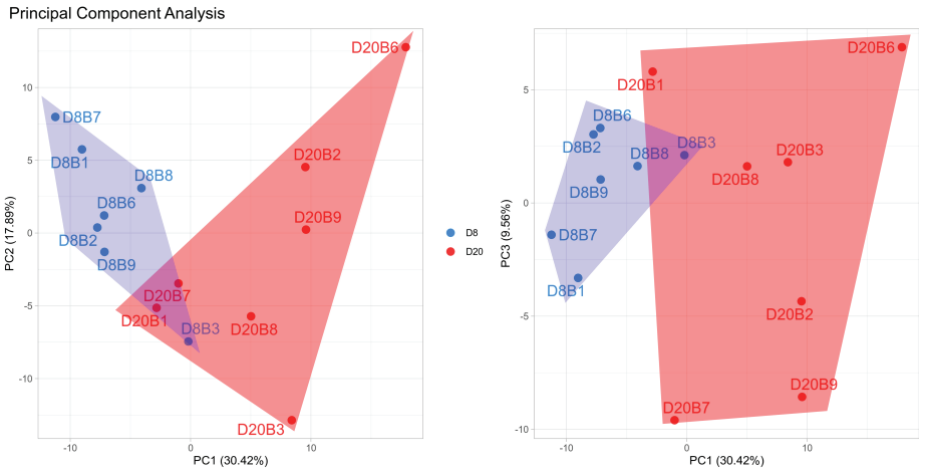

Figure 5-2: PCA of the normalised gene expression counts. There is less variance in the 8dpi vs 20dpi samples as well as some overlap between time points.

We wanted to find specific genes related to high parasitemia without the time factor playing a part, so the D8 samples were divided into two groups based on parasitemia (high or low). We discovered that only the large ribosomal gene exhibited differential expression. The DGE analysis, that is typically used in

this type of studies, posed a challenge due to the need for numerous correlation tests and need to account for potential false negatives. To simplify the analysis, we used the correlation of the PCA values with parasitemia. Then we focused on genes with the highest loading coefficient to the specific principal component. Although this approach weakened the direct connection between parasitemia and specific genes, it allowed us to investigate if certain biological processes, identified through GO-term enrichment analysis, were significantly overrepresented in the genes contributing to the significant principal component. We observed an enrichment of genes related to cell replication and cell movement among those associated with high parasitemia. Conversely, GO terms linked to low parasitemia were related to cellular metabolism. This replicates the results we saw in the time-point DGE analysis.

5.1.3 Shifts in Variation

As seen in the PCA (Figure 5-2), the variation in the gene expression of the D20 samples was significantly higher than in the D8 samples. This was confirmed (p=8e-7) using Feltz and Miller's (1996) asymptotic test for differences in coefficient of variation which we preformed post publication. We speculate on the causes for this increased variation: A) as time progresses, the parasites lifecycles desynchronise between host individuals resulting in different expression profiles, and B) as each bird individual is genetically distinct, its immune response might be different, and if so the parasite will adapt in turn. There is a third factor that is explored more in Paper 5, which is that the genetic composition of parasite populations might change between host individuals over the course of infection, and this change might be different between individual infections depending on whether neutral or natural selection is at play.

5.2 Paper 2

5.2.1 Less variation than expected

The first 9 haplotypes for block 14 of the MSP1 gene in *Plasmodium relictum* were first identified in Hellgren *et al.* (2015) and are named Pr1 to Pr9, and the two new haplotypes in this paper are named Pr10 and Pr11. Of the 75 samples, GRW11 was identified in 11 samples and SGS1 in 64 samples using the *cyt. b*

gene, and 72 of them had the Pr2 haplotype from across the Palearctic region. Of the remaining three samples, Pr3 was found in a GRW11 sample, and the two new haplotypes, Pr10 and Pr11, were found in SGS1 samples originating from adult migratory birds that originate from the Afrotropical and Western Palearctic regions, respectively (Figure 5-3). Pr10 is only one non-synonymous mutation different from Pr1, which is also primarily transmitted in sub-Saharan Africa, and Pr11 is also just differing in one synonymous mutation from Pr2, the primary haplotype in the Western Palearctic. This demonstrates very low diversity in a gene which is under strong pressure from the host's immune defence, especially when one considers these samples were taken from 9 species for the GRW11 samples and 25 species for the SGS1 samples.

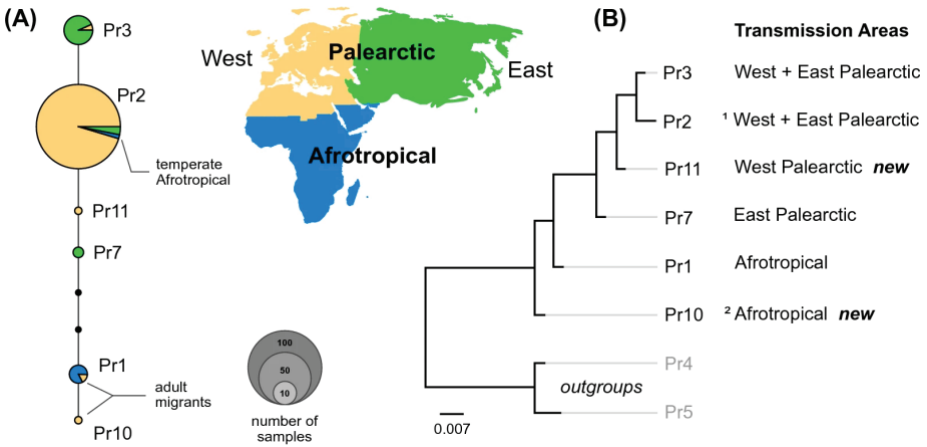

Figure 5-3: **A**: Haplotype network and geographic map of the most common SGS1 and GRW11 haplotypes. **B**: Phylogenetic tree of the same haplotypes highlighting geographic grouped clades.

5.2.2 Minimal interbreeding

SGS1 populations in Africa and Europe do not appear to be interbreeding, as we do not see juveniles or resident bird species in Europe with the Pr1 or Pr10 haplotypes. Pr10 was found in adult migrant birds that winter in Africa, which is the likely source area for transmission of Pr10. Sampling in the Afrotropical region would provide valuable data regarding the genetic composition of *P. relictum* in locally resident birds, and especially confirm the status of SGS1

and GRW11 in the region, as they are not typically thought of as tropical lineages (GRW11 has never been found in Africa). This data confirmed the initial finding in Hellgren (2015) that the Western Palearctic population exhibits very low genetic variation even when including more species and that the division between an African population and a temperate (Euro-Asian) population of SGS1 seems to be valid.

5.3 Paper 3

5.3.1 Coverage

The targeted sequencing capture success varied widely across samples, lineages, and parasitemia levels (Figure 5-4). Generally, the higher the parasitemia and genetic relatedness, the better the coverage. The GRW11 samples had a very high parasitemia and very high coverage, with 99.78 % of the targeted basepairs with at least 5x coverage, but the GRW4 sample only had 24 % coverage at 5x depth, which may be due to very low (1%) parasitemia and greater genetic distance to the reference. Coverage tends to be saturated from > 7 % parasitemia, and parasitemia less than 1 % provides less than 30 % coverage.

Even with less than 30% coverage for low parasitemia samples, we still had enough data to construct haplotype networks for 25 genes, with highlighted genetic diversity within and between lineages within the *P. relictum* morpho-specie. We saw differences within the SGS1 samples, between SGS1 and the reference genome (DONANA05) and with GRW11, the latter thought to be interbreeding with SGS1. The greatest difference was with GRW4, with 663 differences compared to SGS1 over 43Kb (1.5 % divergence). Compared with previous studies that used only one or two markers, much greater phylogenetic resolution is available with just a few more markers, revealing levels of divergence that could separate SGS1 and GRW4 as species with a higher confidence. This species delineation can be confirmed after more sampling to understand within and between lineage divergence.

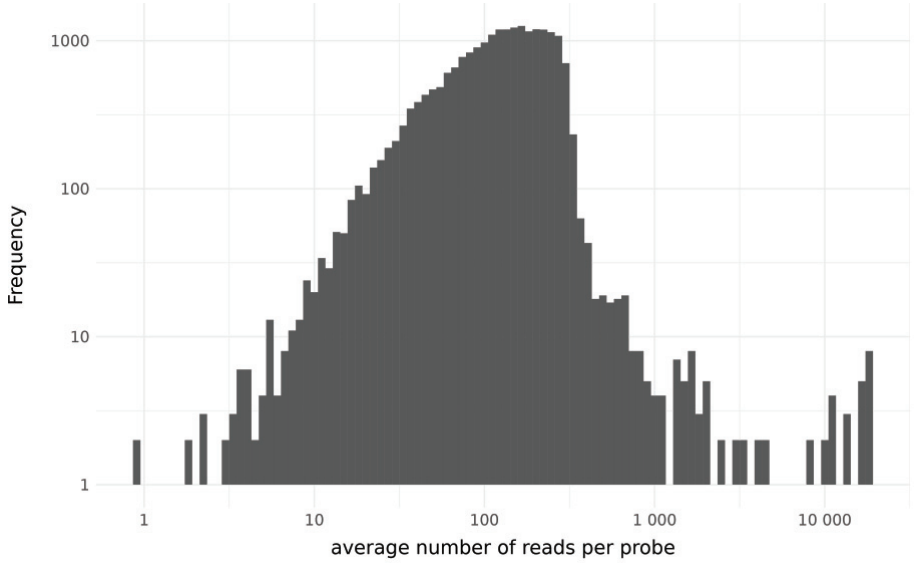

Figure 5-4: Histogram of reads per probed region averaged across all samples. Note that both axes are $\log^{10}$ transformed. Most of the probes were covered by over 15 reads.

5.3.2 Visualisation

The visualisation methods used in this study were highly informative and are recommended for similar studies. They revealed extremely uneven depth of coverage over the targeted regions, with many "spikes" of coverage with over 2000x spanning just a few hundred bases. These regions were found in most chromosomes and samples, including very low parasitemia samples. Most of them were composed of low complexity sequences, therefore the baits may have been overly effective, or the genomic mapping software may have 'overmapped' to those regions, 'stealing' from other paralogous regions in the genome. The probes for these regions are removed from future studies, since neighbouring upstream reads will provide sufficient coverage.

5.4 Paper 4

5.4.1 Clinical outcomes

We measured the body weight, haemoglobin, hematocrit and parasitemia for six infected and six control birds used in the experiment. The statistical analysis of the blood parameters did not reveal large differences between the groups, though I2 did have worse haemoglobin and hematocrit levels compared to their controls. By day 20 after infection, three birds from the I2 group succumbed to the infection, while one bird from the I1 group died on day 8. The three deceased birds from I2 and three birds from I1 were chosen for RNA extraction.

5.4.2 Variation among samples

The clinical outcomes (mortality and parasitemia) were linked with the parasite genotypes for each group. The phylogeny distinctly separates the I1 and I2 groups, and the UpSet plot shows a unique set of 23 SNPs (alternative bases for the same position) that delineate the two groups. However, we also saw genetic variation within the groups, as each sample was genetically unique.

We concluded that genetic variation was present in the donor bird by observing genetic variation in the recipient hosts. This implies that each infection contains a population of parasite haplotypes, but the majority consensus haplotype is what we construct after sequencing the populations in the recipients (see 4.4.2 final paragraph). The majority consensus haplotypes for each group share alternative bases for the same position which outline the differences between the two donor birds.

Each host in each group had a unique majority consensus haplotype profile, with both unique and shared SNPs between the samples (Figure 5-5), which may be due to various factors: a) when blood is transferred from the donor to recipients, several stages of random selection occur, resulting in a variation in the combination of haplotypes received by each recipient; b) then as each host is genetically and phenotypically unique, it will impose different resource limitations and immune pressures on the population, which might select for specific haplotypes; and finally c) it is possible that some of this variation is due to novel mutations.

Sequencing the infected blood from the donors before it is inoculated into the recipients may shed light on these questions, but was not done in the current studies. There is also the fact that we are sequencing mRNA, so we need to consider alternative splicing, mRNA editing, and RNA degradation which will change the sequencing outcome.

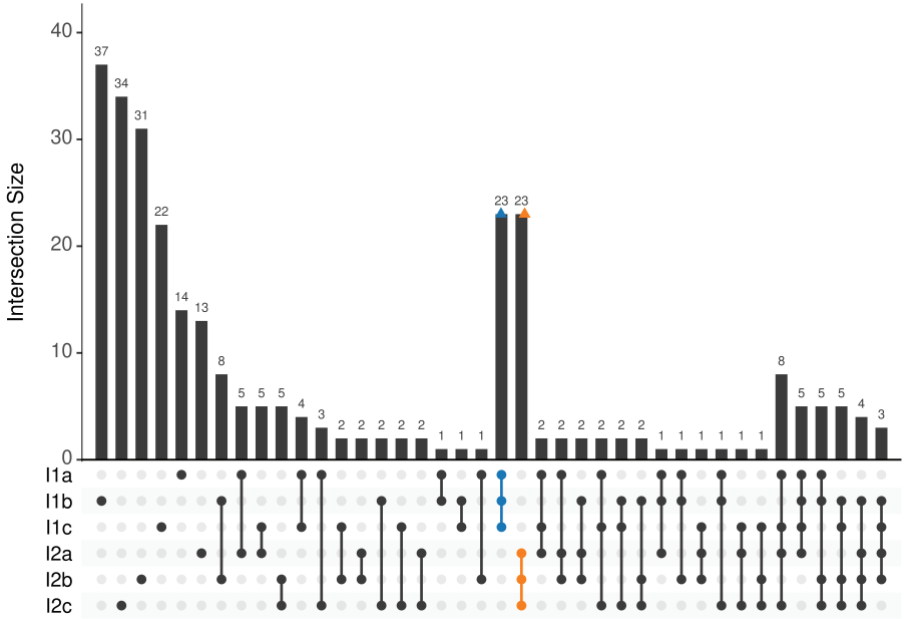

Figure 5-5: UpSet plot of the SNPs present in each sample. The two infection groups share alternative bases for 23 SNPs which delineate them genetically. We also see genetic variation within the infection groups.

The 23 SNPs that separate the two groups genetically were found in 20 genes, however there is no consensus on the relevancy for the functions of those genes and the differences in the phenotypes we observe. We need a larger study with more hosts which also take into account the genotype and phenotype of the hosts to identify specific genes and SNPs which influence parasitemia and host mortality.

5.5 Paper 5

This study allowed us to see how haplotype populations change over the course of an infection, and it drew attention to changes in genes that were only found in samples with unsuppressed infections.

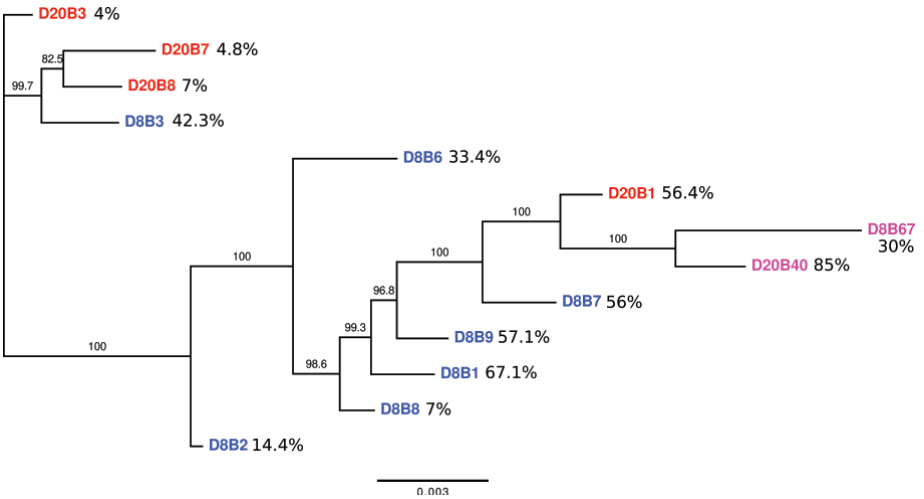

Figure 5-6: Maximum likelihood phylogeny with SH-aLRT results as branch support values. All but one branch has less than 95% confidence. The samples from day 8 and day 20 are labelled with a D8 and D20 and with blue and red respectively. The purple samples were not included in Paper I. The percentages represent parasitemia.

5.5.1 Intersample Variation

All of the recipient birds received their inoculation from the same donor bird and, just as in Paper 4, large genetic variation was observed between our day 8 samples and there were no genetically identical samples. It is evident that the donor bird harbours a population of parasites, and it is unlikely that the composition of the transferred parasites is the same across all the recipients, simply due to random events during the inoculation. However, drift (random), bottlenecks (selection) and mutation (random and selection) during the early replication phase could also play a role in the variation seen between samples.

5.5.2 Time and Parasitemia

These same factors may explain why the haplotypes were not stable across time, as can be seen in Figure 5-6, where no haplotype remained the same between 8dpi and 20dpi. Since every host is genetically unique, they will exert distinct immune pressures, leading to the emergence of various "escape haplotypes." This variation in escape haplotypes will become more pronounced as the infection persists over time.

We only had four host individuals that were sampled at both time points, which makes it harder to analyse the haplotype dynamics over time. However, the phylogenetic tree topology was well supported and it depicted that B7 and B8 moved across the tree between time points. Our attention was drawn to two main clades: the clade at the base of the tree (including the samples D20B3, D20B7, D20B8, D8B3) and the clade that forms the rest of the tree. The samples that had the highest parasitemia over time (D20B1, D20B40, D8B67, see Paper 5: Figure 1) all cluster together with the longest branches. B67 is remarkable as it died on day 19 (the only bird to die during the experiment, with 90% of its red blood cells infected), and the infection in B40 flared up post day 24, unique among the samples.

These two ‘susceptible’ samples form their own node on the tree, and share 59 exclusive (Paper 5: Figure 3) SNPs across 17 genes. Twelve of the 17 genes are part of large gene families involved in cell invasion, immune evasion, and intraerythrocytic parasite growth, with the final five genes being unrelated to these functions or are unannotated. It is difficult to determine the direction of causality in these two hosts. It may be that the successful haplotypes would have been just as virulent in other hosts, or it may be that these two susceptible hosts were unfortunate in their specific interactions with these haplotypes and allowed for unsuppressed infections making these specific haplotypes rise to become the major haplotype in the infection. It is possible that the dominant haplotypes in B40 and B67 were present at low levels in the other samples, but drift or selection prevented their growth. Once again, larger studies are required to pin down parasite-host interactions on the gene level to understand these unsuppressed infections.

5.5.3 Linking haplotypes and transcription

When we look back to the gene count dendrogram and PCA in the RNA-seq expression analysis in Paper 1, we see a similar clustering in the phylogeny, where the B8 samples clustered with the D20 samples and the D20B1 clustered

with the D8 samples (Figure 5-7). This concordance between the phylogenetic and gene expression data could be attributed to the genetic determination of the parasite's expression in response to the host, resulting in similarities. The phylogenetic data, in turn, reflects this similarity within the clades on the tree. The clustering of gene count data is likely influenced by parasitemia, but parasitemia will be influenced by the parasite's genetics and host response.

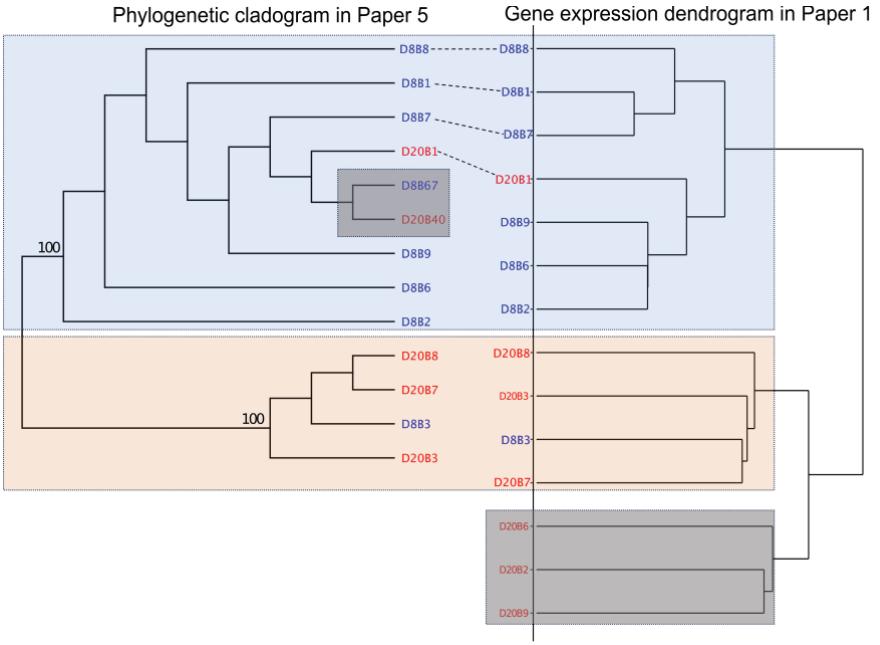

Figure 5-7: A comparison of the clustering of the phylogenetic cladogram and gene expression dendrogram from Paper 1. There is strong congruence in the general structure of the trees. The phylogenetic tree from Figure 5-6 was converted to a cladogram with rerooting to make the comparison easier. The grey boxes cover samples that are not present in both studies. The blue and red shaded boxes represent the 2 main clusters of congruence. The blue and red sample tips are from 8 and 20 days post infection respectively.

6 Conclusions

Genetic variation abounds in *Plasmodium relictum* as soon as one looks beyond the surface level of just one or two genes. In terms of gene expression, patterns of variation can be seen as the parasites adapt to their respective hosts, increasing over time and in complexity. The European populations of SGS1 and GRW11 appear to be monomorphic when analysing an otherwise variable block of *msp1,* which is surprising given the function of this gene and its diversity in human malaria. However, targeted exome and mRNA sequencing revealed significant variation. GRW4 is much more dissimilar compared to SGS1 and GRW11, which has been suspected given that they are largely allopatric populations, but is now confirmed in Paper 3 thanks to the much larger genomic data set. Paper 3 also hinted at within lineage genetic variation, which we were surprised by in Papers 4 and 5.

The identification of a population of haplotypes within an infection, as demonstrated in Paper 4, is a critical finding. It suggests that an infection is not just a monolithic entity but rather a dynamic and diverse collection of haplotypes. This finding has far-reaching implications for understanding the dynamics of infections and their potential for adaptation and evolution.

Paper 4 revealed large, fixed differences between infection groups which also had differences in mortality, but also genetic diversity within the groups. The findings in Paper 5 provide a crucial dimension to the previous research, highlighting the dynamic nature of the infection population over time. We observed how some haplotypes can result in unsuppressed and mortal infections, which highlighted specific mutations and genes. Furthermore, the observation that these changes are influenced by the host adds a layer of complexity and significance to the understanding of the parasite's behaviour and interaction with its hosts.

6.1 History and Caveats

In fact, I was surprised by the level of genetic diversity within a single infection. The very first attempt I made to quantify the variation followed the standard GATK workflows for genomic and RNA-seq genotyping, but I assumed the VCF files I was producing were garbage given the large number

of SNPs after filtering. After using the less traditional methods shown here (but methods I had a better understanding of), I look back and wonder if I threw the baby out with the bathwater. However, the standard workflows usually require high quality SNP databases used in model organisms, which are not available here, and they assume diploid genomes.

By creating majority consensus haplotypes for each sample, I reject all the rare SNPs, and rely on BCFtools to call SNPs given an unknown ploidy. To get reliable haplotypes within an infection, we require much longer sequencing reads, and high parasitemia with deep sequencing to detect the rarer haplotypes. In preprint, Niare *et al* (2023) describe a variant calling protocol for *P. falciparum* that uses an optimised GATK4 pipeline and PacBio assemblies applied to Illumina short read sequences, resulting in superior sensitivity. Longer reads provide the additional benefit that SNPs across a chromosome can be linked, which allows for a degree of phasing haplotypes (Laver *et al.*, 2016; Sakamoto, Sereewattanawoot and Suzuki, 2019).

Another caveat is the quality of the reference genome. Many of the genes with many SNPs were found in the non-chromosomal contigs (so called archive chromosomes) which could not be assembled onto the main chromosomes given the low complexity nature of the AT rich genome. In order to get sufficient parasite DNA, over 100 oocysts from 50 mosquito midguts were harvested (Böhme *et al.*, 2018), which would result in multiple haplotypes if the assembly process did not flatten them out. A side effect of this process could be the increase in the number of copies of highly polymorphic genes, like some of the gene families highlighted in this thesis. This reduces confidence in the annotations and creates doubts about whether we are looking at orthologues, paralogues or highly variable genes.

6.2 The Future of Avian Malaria Research

A follow up study for Paper 5 should use more birds and more time points, and sequence the inoculum before it is inoculated into the birds, which would help distinguish the effects of bottlenecks, drift, and mutation in the evolution of an infection. Dual-RNAseq transcriptomes can be used to understand the gene-gene interactions between the parasite and the host, which can still be attempted with the data in Paper 4. One can add complexity to the experiments by cross infecting hosts from and to different species, and observing which haplotypes are transmitted between and with-in host species, and their subsequent effects on virulence.

It is now vital to understand the effect of the mosquito, which is a source of variation through sexual reproduction and a bottleneck in transmission, because understanding the variation in the population of parasites in the vector will help us understand parasite evolution.

The collective impact of these findings goes beyond the specific infection system studied and holds significant implications for medical research and evolutionary biology. Understanding the genetic diversity within infections and how they evolve over time can inform better treatment strategies, vaccine development, and a deeper understanding of host-pathogen interactions. The concept of within lineage genetic variation and the population dynamics of infections open up exciting avenues for future studies, promising new discoveries and advancements in our understanding of infectious diseases.

7 References

Aavani, P. and Rice, S.H. (2022) ‘When sexual selection in hosts benefits parasites’, *Infectious Disease Modelling*, 7(3), pp. 561–570. Available at: https://doi.org/10.1016/J.IDM.2022.08.001.

Anders, S., Pyl, P.T. and Huber, W. (2015) ‘HTSeq—a Python framework to work with high-throughput sequencing data’, *Bioinformatics*, 31(2), p. 166. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTU638.

Andrews, S. (2010) ‘FastQC: A Quality Control Tool for High Throughput Sequence Data’. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Asghar, M. *et al.* (2015) ‘Hidden costs of infection: Chronic malaria accelerates telomere degradation and senescence in wild birds’, *Science*, 347(6220), pp. 436–438. Available at: https://doi.org/10.1126/SCIENCE.1261121/SUPPL_FILE/ASGHAR.SM.PDF.

Atkinson, C.T. and LaPointe, D.A. (2009) ‘Introduced Avian Diseases, Climate Change, and the Future of Hawaiian Honeycreepers’, *Journal of Avian Medicine and Surgery*, 23(1), pp. 53–63. Available at: https://doi.org/10.1647/2008-059.1.

Avise, J.C. (1994) *Molecular Markers, Natural History and Evolution*, *Molecular Markers, Natural History and Evolution*. Boston, MA: Springer US. Available at: https://doi.org/10.1007/978-1-4615-2381-9.

Awadalla, P. (2003) ‘The evolutionary genomics of pathogen recombination’, *Nature Reviews Genetics*. Nat Rev Genet, pp. 50–60. Available at: https://doi.org/10.1038/nrg964.

Aželytė, J. *et al.* (2022) ‘A comparative analysis of the dynamics of Plasmodium relictum (GRW4) development in the blood during single and co-infections’, *Acta Tropica*, 226, p. 106247. Available at: https://doi.org/10.1016/J.ACTATROPICA.2021.106247.

van Baalen, M. and Sabelis, M.W. (1995) ‘The dynamics of multiple infection and the evolution of virulence’, *American Naturalist*, 146(6), pp. 881–910. Available at: https://doi.org/10.1086/285830.

Beadell, J.S. *et al.* (2006) 'Global phylogeographic limits of Hawaii's avian malaria', *Proceedings of the Royal Society B: Biological Sciences*, 273(1604), pp. 2935–2944. Available at: https://doi.org/10.1098/RSPB.2006.3671.

Ben-harel, S. (1923) 'Studies of bird malaria in relation to the mechanism of relapse', *American Journal of Epidemiology*, 3(6), pp. 652–685. Available at: https://doi.org/10.1093/OXFORDJOURNALS.AJE.A118961/2/3-6-652.PDF.GIF.

Bensch, S., Hellgren, O. and Pérez-Tris, J. (2009) 'MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages', *Molecular Ecology Resources*, 9(5), pp. 1353–1358. Available at: https://doi.org/10.1111/j.1755-0998.2009.02692.x.

Berngruber, T.W. *et al.* (2013) 'Evolution of Virulence in Emerging Epidemics', *PLoS Pathogens*. Edited by F. Balloux, 9(3), p. e1003209. Available at: https://doi.org/10.1371/journal.ppat.1003209.

Berngruber, T.W., Lion, S. and Gandon, S. (2015) 'Spatial Structure, Transmission Modes and the Evolution of Viral Exploitation Strategies', *PLOS Pathogens*. Edited by S.P. Brown, 11(4), p. e1004810. Available at: https://doi.org/10.1371/journal.ppat.1004810.

Böhme, U. *et al.* (2018) 'Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution in birds and mammals', *Genome Research*, 28(4), pp. 547–560. Available at: https://doi.org/10.1101/gr.218123.116.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTU170.

Bopp, S.E.R. *et al.* (2013) 'Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families', *PLOS Genetics*, 9(2), p. e1003293. Available at: https://doi.org/10.1371/JOURNAL.PGEN.1003293.

Brown, S.P., Cornforth, D.M. and Mideo, N. (2012) 'Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control', *Trends in Microbiology*, 20(7), pp. 336–342. Available at: https://doi.org/10.1016/j.tim.2012.04.005.

Camponovo, F. *et al.* (2021) ‘Mechanistic within-host models of the asexual Plasmodium falciparum infection: a review and analytical assessment’, *Malaria Journal*, 20(1), pp. 1–22. Available at: https://doi.org/10.1186/S12936-021-03813-Z/FIGURES/6.

Chagas, C.R.F., Harl, J. and Valkiūnas, G. (2021) ‘Co-infections of Plasmodium relictum lineages pSGS1 and pGRW04 are readily distinguishable by broadly used PCR-based protocols, with remarks on global distribution of these malaria parasites’, *Acta Tropica*, 217, p. 105860. Available at: https://doi.org/10.1016/J.ACTATROPICA.2021.105860.

Chua, C.L.L. *et al.* (2021) ‘Poor Birth Outcomes in Malaria in Pregnancy: Recent Insights Into Mechanisms and Prevention Approaches’, *Frontiers in Immunology*, 12, p. 621382. Available at: https://doi.org/10.3389/FIMMU.2021.621382.

Conway, J.R., Lex, A. and Gehlenborg, N. (2017) ‘UpSetR: An R package for the visualization of intersecting sets and their properties’, *Bioinformatics*, 33(18). Available at: https://doi.org/10.1093/bioinformatics/btx364.

Cowell, A.N. *et al.* (2017) ‘Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of Plasmodium vivax from unprocessed clinical samples’, *mBio*, 8(1). Available at: https://doi.org/10.1128/MBIO.02257-16/SUPPL_FILE/MBO001173166ST2.DOCX.

Cowman, A.F. *et al.* (2016) ‘Malaria: Biology and Disease’, *Cell*, 167(3), pp. 610–624. Available at: https://doi.org/10.1016/J.CELL.2016.07.055.

Cressler, C.E. *et al.* (2016) ‘The adaptive evolution of virulence: A review of theoretical predictions and empirical tests’, *Parasitology*. Cambridge University Press, pp. 915–930. Available at: https://doi.org/10.1017/S003118201500092X.

Danecek, P. *et al.* (2021) ‘Twelve years of SAMtools and BCFtools’, *GigaScience*, 10(2). Available at: https://doi.org/10.1093/GIGASCIENCE/GIAB008.

Despommier, D.D. (2016) *People, Parasites, and Plowshares*, *People, Parasites, and Plowshares*. Columbia University Press. Available at: https://doi.org/10.7312/desp16194.

Devictor, V. *et al.* (2010) ‘Defining and measuring ecological specialization’, *Journal of Applied Ecology*, 47(1), pp. 15–25. Available at: https://doi.org/10.1111/j.1365-2664.2009.01744.x.

Dobin, A. *et al.* (2013) ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics (Oxford, England)*, 29(1), pp. 15–21. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTS635.

Edgar, R.C. (2004) ‘MUSCLE: A multiple sequence alignment method with reduced time and space complexity’, *BMC Bioinformatics*, 5(1), pp. 1–19. Available at: https://doi.org/10.1186/1471-2105-5-113/FIGURES/16.

Ewald, P.W. (1983) ‘Host-parasite Relations, Vectors, and the Evolution of Disease Severity’, *Ann. Rev. Ecol. Syst*, 14, pp. 465–85. Available at: www.annualreviews.org (Accessed: 30 August 2020).

Ewald, P.W. (1994) *Evolution of infectious disease*. Edited by F.A.O. of the UN. Oxford University Press.

Ewels, P. *et al.* (2016) ‘MultiQC: Summarize analysis results for multiple tools and samples in a single report’, *Bioinformatics*, 32(19), pp. 3047–3048. Available at: https://doi.org/10.1093/bioinformatics/btw354.

Fecchio, A., Wells, K., Bell, J.A., Tkach, V. V., *et al.* (2019) ‘Climate variation influences host specificity in avian malaria parasites’, *Ecology Letters*, 22(3), pp. 547–557. Available at: https://doi.org/10.1111/ELE.13215.

Fecchio, A., Wells, K., Bell, J.A., Tkach, V. v., *et al.* (2019) ‘Climate variation influences host specificity in avian malaria parasites’, *Ecology Letters*. Edited by P. Thrall, 22(3), pp. 547–557. Available at: https://doi.org/10.1111/ele.13215.

Feltz, C.J. and Miller, G.E. (1996) ‘An Asymptotic Test for the Equality of Coefficients of Variation from k Populations’, *STATISTICS IN MEDICINE*, 15, pp. 647–658. Available at: https://doi.org/10.1002/(SICI)1097-0258(19960330)15:6.

Frank, S.A. (1996) ‘Models of Parasite Virulence’, *The Quarterly Review of Biology*, 71(1), pp. 37–78. Available at: https://doi.org/10.1086/419267.

Futuyma, D.J. and Moreno, G. (1988) *The Evolution of Ecological Specialization*, *Ann. Rev. Ecol. Syst*. Available at: www.annualreviews.org (Accessed: 31 August 2020).

Garamszegi, L.Z. (2011) ‘Climate change increases the risk of malaria in birds’, *Global Change Biology*, 17(5), pp. 1751–1759. Available at: https://doi.org/10.1111/J.1365-2486.2010.02346.X.

Garamszegi, L.Z. and Møller, A.P. (2007) ‘Prevalence of avian influenza and host ecology’, *Proceedings of the Royal Society B: Biological Sciences*, 274(1621), pp. 2003–2012. Available at: https://doi.org/10.1098/RSPB.2007.0124.

Garcia, C.R.S., Markus, R.P. and Madeira, L. (2001) ‘Tertian and Quartan Fevers: Temporal Regulation in Malarial Infection’, *JOURNAL OF BIOLOGICAL RHYTHMS* [Preprint].

Griffin, A.S. and West, S.A. (2002) ‘Kin selection: fact and fiction’, *Trends in Ecology & Evolution*, 17(1), pp. 15–21. Available at: https://doi.org/10.1016/S0169-5347(01)02355-2.

Gupta, P. *et al.* (2020) ‘Host phylogeny matters: Examining sources of variation in infection risk by blood parasites across a tropical montane bird community in India’, *Parasites and Vectors*, 13(1), pp. 1–13. Available at: https://doi.org/10.1186/S13071-020-04404-8/FIGURES/4.

Haine, E.R. (2008) ‘Symbiont-mediated protection’, *Proceedings of the Royal Society B: Biological Sciences*, 275(1633), pp. 353–361. Available at: https://doi.org/10.1098/rspb.2007.1211.

Harrison, F. *et al.* (2006) ‘Cooperation and virulence in acute Pseudomonas aeruginosainfections’, *BMC Biology*, 4(1), p. 21. Available at: https://doi.org/10.1186/1741-7007-4-21.

Hellgren, O., Waldenström, J., *et al.* (2007) ‘Detecting shifts of transmission areas in avian blood parasites — a phylogenetic approach’, *Molecular Ecology*, 16(6), pp. 1281–1290. Available at: https://doi.org/10.1111/J.1365-294X.2007.03227.X.

Hellgren, O., Križanauskiene, A., *et al.* (2007) ‘Diversity and Phylogeny of Mitochondrial Cytochrome B Lineages From Six Morphospecies of Avian Haemoproteus (Haemosporida: Haemoproteidae)’, *Journal of Parasitology*, 93(4), pp. 889–896. Available at: https://doi.org/10.1645/GE-1051R1.1.

Hellgren, O. *et al.* (2015) ‘Global phylogeography of the avian malaria pathogen Plasmodium relictum based on MSP1 allelic diversity’, *Ecography*, 38(8), pp. 842–850. Available at: https://doi.org/10.1111/ecog.01158.

Herre, E.A. (1993) ‘Population Structure and the Evolution of Virulence in Nematode Parasites of Fig Wasps’, *Science*, 259(5100), pp. 1442–1445. Available at: https://doi.org/10.1126/science.259.5100.1442.

Hewitt, R. (1940) ‘Bird malaria [monograph]’, *Am J Hyg Monogr*, 15, p. 65.

Hoberg, E.P. and Brooks, D.R. (2015) ‘Evolution in action: climate change, biodiversity dynamics and emerging infectious disease’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1665), p. 20130553. Available at: https://doi.org/10.1098/rstb.2013.0553.

Huang, X. *et al.* (2019) ‘Within-Lineage Divergence of Avian Haemosporidians: A Case Study to Reveal the Origin of a Widespread Haemoproteus Parasite’, *Journal of Parasitology*, 105(3), p. 414. Available at: https://doi.org/10.1645/18-112.

Humphries, M.B., Stacy, M.T. and Ricklefs, R.E. (2019) ‘Population structure of avian malaria parasites’, *Ecology and Evolution*, 9(13), pp. 7741–7751. Available at: https://doi.org/10.1002/ece3.5356.

Ilgunas, M. *et al.* (2016) ‘Mortality and pathology in birds due to Plasmodium (Giovannolaia) homocircumflexum infection, with emphasis on the exoerythrocytic development of avian malaria parasites’, *Malaria Journal*, 15(1), pp. 1–11. Available at: https://doi.org/10.1186/S12936-016-1310-X/FIGURES/5.

Kalbskopf, V. *et al.* (2021) ‘Shifts in gene expression variability in the blood-stage of Plasmodium relictum’, *Gene*, 792, p. 145723. Available at: https://doi.org/10.1016/j.gene.2021.145723.

Kalbskopf, V. *et al.* (no date) ‘Linking infection outcome to genomic variation within the avian malaria lineage SGS1, Plasmodium relictum using experimental infections and RNA-sequencing.’, *Manuscript* [Preprint].

Kalyaanamoorthy, S. *et al.* (2017) ‘ModelFinder: Fast model selection for accurate phylogenetic estimates’, *Nature Methods*, 14(6). Available at: https://doi.org/10.1038/nmeth.4285.

Kim, D. *et al.* (2019) ‘Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype’, *Nature Biotechnology*, 37(8), pp. 907–915. Available at: https://doi.org/10.1038/s41587-019-0201-4.

Knowles, S.C.L., Palinauskas, V. and Sheldon, B.C. (2010) ‘Chronic malaria infections increase family inequalities and reduce parental fitness:

experimental evidence from a wild bird population', *Journal of Evolutionary Biology*, 23(3), pp. 557–569. Available at: https://doi.org/10.1111/J.1420-9101.2009.01920.X.

Kuris, A.M. *et al.* (2008) 'Ecosystem energetic implications of parasite and free-living biomass in three estuaries', *Nature 2008 454:7203*, 454(7203), pp. 515–518. Available at: https://doi.org/10.1038/NATURE06970.

Lauron, E.J. *et al.* (2014) 'Transcriptome sequencing and analysis of Plasmodium gallinaceum reveals polymorphisms and selection on the apical membrane antigen-1', *Malaria Journal*, 13(1), pp. 1–14. Available at: https://doi.org/10.1186/1475-2875-13-382/FIGURES/6.

Lauron, E.J. *et al.* (2015) 'De novo assembly and transcriptome analysis of Plasmodium gallinaceum identifies the Rh5 interacting protein (ripr), and reveals a lack of EBL and RH gene family diversification', *Malaria Journal*, 14(1), pp. 1–10. Available at: https://doi.org/10.1186/S12936-015-0814-0/FIGURES/3.

Laver, T.W. *et al.* (2016) 'Pitfalls of haplotype phasing from amplicon-based long-read sequencing', *Scientific Reports 2016 6:1*, 6(1), pp. 1–6. Available at: https://doi.org/10.1038/SREP21746.

Laveran, A. (1880) 'Note sur un nouveau parasite trouvé dans le sang de plusieurs malades atteints de fièvre palustre.(Première note) Bull', *Acad. Med*, pp. 1235–1236.

Leigh, J.W. and Bryant, D. (2015) 'popart: full-feature software for haplotype network construction', *Methods in Ecology and Evolution*, 6(9), pp. 1110–1116. Available at: https://doi.org/10.1111/2041-210X.12410.

LEVINE, N.D. (1988) 'Progress in Taxonomy of the Apicomplexan Protozoa', *The Journal of Protozoology*, 35(4), pp. 518–520. Available at: https://doi.org/10.1111/J.1550-7408.1988.TB04141.X.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTP352.

Lim, L. and McFadden, G.I. (2010) 'The evolution, metabolism and functions of the apicoplast', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1541), p. 749. Available at: https://doi.org/10.1098/RSTB.2009.0273.

Lipsitch, M. *et al.* (1995) 'The population dynamics of vertically and horizontally transmitted parasites', *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 260(1359), pp. 321–327. Available at: https://doi.org/10.1098/rspb.1995.0099.

Little, T.J., Watt, K. and Ebert, D. (2006) 'Parasite-host Specificity: Experimental Studies on the Basis of Parasite Adaptation', *Evolution*, 60(1), pp. 31–38. Available at: https://doi.org/10.1111/J.0014-3820.2006.TB01079.X.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Lowe, S. *et al.* (2000) *100 of the world's worst invasive alien species: a selection from the global invasive species database*. Invasive Species Specialist Group Auckland.

Malaria deaths by age, World, 1990 to 2019 (no date). Available at: https://ourworldindata.org/grapher/malaria-deaths-by-age (Accessed: 8 May 2023).

Manwell, R.D. and Goldstein, F. (1938) 'Life History and Immunity Studies of the Avian Malaria Parasite, Plasmodium circumflexum.', *https://doi-org.ludwig.lub.lu.se/10.3181/00379727-39-10226P*, 39(3), pp. 426–428. Available at: https://doi.org/10.3181/00379727-39-10226P.

Marzal, A. *et al.* (2015) 'Invasive avian malaria as an emerging parasitic disease in native birds of Peru', *Biological Invasions*, 17(1), pp. 39–45. Available at: https://doi.org/10.1007/S10530-014-0718-X/TABLES/1.

de Meeûs, T., Lehmann, L. and Balloux, F. (2006) 'Molecular epidemiology of clonal diploids: A quick overview and a short DIY (do it yourself) notice', *Infection, Genetics and Evolution*, 6(2), pp. 163–170. Available at: https://doi.org/10.1016/j.meegid.2005.02.004.

Meister, S.L. *et al.* (2021) 'Fatal avian malaria in captive Atlantic puffins (Fratercula arctica) in Switzerland', *International Journal for Parasitology: Parasites and Wildlife*, 14, pp. 97–106. Available at: https://doi.org/10.1016/j.ijppaw.2020.12.007.

Michod, R.E., Bernstein, H. and Nedelcu, A.M. (2008) 'Adaptive value of sex in microbial pathogens', *Infection, Genetics and Evolution*, 8(3), pp. 267–285. Available at: https://doi.org/10.1016/j.meegid.2008.01.002.

Murray, L. *et al.* (2017) ‘Multiplication rate variation in the human malaria parasite Plasmodium falciparum’, *Scientific Reports 2017 7:1*, 7(1), pp. 1–8. Available at: https://doi.org/10.1038/s41598-017-06295-9.

Mwingira, F. *et al.* (2011) ‘Plasmodium falciparum msp1, msp2 and glurp allele frequency and diversity in sub-Saharan Africa’, *Malaria Journal*, 10. Available at: https://doi.org/10.1186/1475-2875-10-79.

Neafsey, D.E., Taylor, A.R. and MacInnis, B.L. (2021) ‘Advances and opportunities in malaria population genomics’, *Nature Reviews. Genetics*, 22(8), p. 502. Available at: https://doi.org/10.1038/S41576-021-00349-5.

Nguyen, L.T. *et al.* (2015) ‘IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies’, *Molecular Biology and Evolution*, 32(1). Available at: https://doi.org/10.1093/molbev/msu300.

Niaré, K., Greenhouse, B. and Bailey, J.A. (2023) ‘An Optimized GATK4 Pipeline for Plasmodium falciparum Whole Genome Sequencing Variant Calling and Analysis’, *Research Square* [Preprint]. Available at: https://doi.org/10.21203/RS.3.RS-2561857/V1.

Nowak, M.A. and May, R.M. (1994) ‘Superinfection and the evolution of parasite virulence’, *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342), pp. 81–89. Available at: https://doi.org/10.1098/rspb.1994.0012.

Nylin, S. *et al.* (2018) ‘Embracing Colonizations: A New Paradigm for Species Association Dynamics’, *Trends in Ecology & Evolution*, 33(1), pp. 4–14. Available at: https://doi.org/10.1016/j.tree.2017.10.005.

O’Connor, E.A. *et al.* (2019) ‘Avian MHC Evolution in the Era of Genomics: Phase 1.0’, *Cells 2019, Vol. 8, Page 1152*, 8(10), p. 1152. Available at: https://doi.org/10.3390/CELLS8101152.

Oyola, S.O. *et al.* (2016) ‘Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification’, *Malaria Journal*, 15(1), pp. 1–12. Available at: https://doi.org/10.1186/S12936-016-1641-7/FIGURES/9.

Palinauskas, V. *et al.* (2008) ‘Plasmodium relictum (lineage P-SGS1): Effects on experimentally infected passerine birds’, *Experimental Parasitology*, 120(4), pp. 372–380. Available at: https://doi.org/10.1016/J.EXPPARA.2008.09.001.

Palinauskas, V. *et al.* (2018) ‘Different paths – the same virulence: experimental study on avian single and co-infections with Plasmodium relictum and Plasmodium elongatum’, *International Journal for Parasitology*, 48(14), pp. 1089–1096. Available at: https://doi.org/10.1016/J.IJPARA.2018.08.003.

Palinauskas, V. *et al.* (2020) ‘Dynamics of blood stage and sporozoite-induced malarial infections in experimentally infected passerines’, *International Journal for Parasitology*, 50(13), pp. 1057–1065. Available at: https://doi.org/10.1016/J.IJPARA.2020.05.015.

Pérez-Rubio, P., Lottaz, C. and Engelmann, J.C. (2019) ‘FastqPuri: High-performance preprocessing of RNA-seq data’, *BMC Bioinformatics*, 20(1), pp. 1–11. Available at: https://doi.org/10.1186/s12859-019-2799-0.

Phillips, M.A. *et al.* (2017) ‘Malaria’, *Nature Reviews Disease Primers 2017 3:1*, 3(1), pp. 1–24. Available at: https://doi.org/10.1038/nrdp.2017.50.

Picard Tools - By Broad Institute (no date). Available at: http://broadinstitute.github.io/picard/ (Accessed: 8 March 2023).

Poulin, R. (2010) ‘Network analysis shining light on parasite ecology and diversity’, *Trends in Parasitology*, 26(10), pp. 492–498. Available at: https://doi.org/10.1016/j.pt.2010.05.008.

Poulin, R. (2011) *Evolutionary ecology of parasites*, *Evolutionary Ecology of Parasites: (Second Edition)*. Available at: https://doi.org/10.2307/3285737.

Quinlan, A.R. and Hall, I.M. (2010) ‘BEDTools: a flexible suite of utilities for comparing genomic features’, *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

Reed, K.D. *et al.* (2003) ‘Birds, Migration and Emerging Zoonoses: West Nile Virus, Lyme Disease, Influenza A and Enteropathogens’, *Clinical Medicine and Research*, 1(1), p. 5. Available at: https://doi.org/10.3121/CMR.1.1.5.

Ricklefs, R.E. *et al.* (2017) ‘Avian migration and the distribution of malaria parasites in New World passerine birds’, *Journal of Biogeography*, 44(5), pp. 1113–1123. Available at: https://doi.org/10.1111/JBI.12928.

Riper, C. van *et al.* (1986) ‘The Epizootiology and Ecological Significance of Malaria in Hawaiian Land Birds’, *Ecological Monographs*, 56(4), pp. 327–344. Available at: https://doi.org/10.2307/1942550.

Rivero, A. and Gandon, S. (2018) ‘Evolutionary Ecology of Avian Malaria: Past to Present’, *Trends in Parasitology*, 34(8), pp. 712–726. Available at: https://doi.org/10.1016/J.PT.2018.06.002.

Rivero-de Aguilar, J. *et al.* (2016) ‘MHC-I provides both quantitative resistance and susceptibility to blood parasites in blue tits in the wild’, *Journal of Avian Biology*, 47(5), pp. 669–677. Available at: https://doi.org/10.1111/JAV.00830.

Roser, M. (no date) *Malaria: One of the leading causes of child deaths, but progress is possible and you can contribute to it - Our World in Data*. Available at: https://ourworldindata.org/malaria-introduction (Accessed: 8 May 2023).

Ross, H.J. (1902) *Letters from the East*. JM Dent & Company.

Sakamoto, Y., Sereewattanawoot, S. and Suzuki, A. (2019) ‘A new era of long-read sequencing for cancer genomics’, *Journal of Human Genetics 2019 65:1*, 65(1), pp. 3–10. Available at: https://doi.org/10.1038/S10038-019-0658-5.

Sedlazeck, F.J., Rescheneder, P. and Von Haeseler, A. (2013) ‘Sequence analysis NextGenMap: fast and accurate read mapping in highly polymorphic genomes’, 29(21), pp. 2790–2791. Available at: https://doi.org/10.1093/bioinformatics/btt468.

Smith, J.M. *et al.* (1993) ‘How clonal are bacteria?’, *Proceedings of the National Academy of Sciences of the United States of America*, 90(10), pp. 4384–4388. Available at: https://doi.org/10.1073/pnas.90.10.4384.

Somé, A.F. *et al.* (2018) ‘Plasmodium falciparum msp1 and msp2 genetic diversity and allele frequencies in parasites isolated from symptomatic malaria patients in Bobo-Dioulasso, Burkina Faso’, *Parasites and Vectors*, 11(1), pp. 1–8. Available at: https://doi.org/10.1186/S13071-018-2895-4/FIGURES/1.

Su, Y., Ruan, S. and Wei, J. (2011) ‘Periodicity and synchronization in blood-stage malaria infection’, *Journal of Mathematical Biology*, 63(3), pp. 557–574. Available at: https://doi.org/10.1007/S00285-010-0381-5/METRICS.

Taylor, A.R. *et al.* (2019) ‘Estimating Relatedness Between Malaria Parasites’, *Genetics*, 212(4), p. 1337. Available at: https://doi.org/10.1534/GENETICS.119.302120.

Theodosopoulos, A.N. *et al.* (2021) 'A highly invasive malaria parasite has expanded its range to non-migratory birds in North America', *Biology Letters*, 17(9). Available at: https://doi.org/10.1098/RSBL.2021.0271.

Thrall, P.H. *et al.* (2007) 'Coevolution of symbiotic mutualists and parasites in a community context', *Trends in Ecology & Evolution*, 22(3), pp. 120–126. Available at: https://doi.org/10.1016/j.tree.2006.11.007.

Tibayrenc, M. (1995) 'Population Genetics of Parasitic Protozoa and other Microorganisms', in *Advances in Parasitology*, pp. 47–115. Available at: https://doi.org/10.1016/S0065-308X(08)60490-X.

Tibayrenc, M. and Ayala, F.J. (2002) 'The clonal theory of parasitic protozoa: 12 years on', *Trends in Parasitology*. Trends Parasitol, pp. 405–410. Available at: https://doi.org/10.1016/S1471-4922(02)02357-7.

Tibayrenc, M. and Ayala, F.J. (2012) 'Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa', *Proceedings of the National Academy of Sciences*, 109(48), pp. E3305–E3313. Available at: https://doi.org/10.1073/pnas.1212452109.

Turner, P.E. and Chao, L. (1999) 'Prisoner's dilemma in an RNA virus', *Nature*, 398(6726), pp. 441–443. Available at: https://doi.org/10.1038/18913.

Videvall, E. *et al.* (2017) 'The transcriptome of the avian malaria parasite Plasmodium ashfordi displays host-specific gene expression', *Molecular Ecology*, 26(11), pp. 2939–2958. Available at: https://doi.org/10.1111/mec.14085.

Videvall, E. (2019) 'Genomic Advances in Avian Malaria Research', *Trends in Parasitology*, 35(3). Available at: https://doi.org/10.1016/j.pt.2018.12.005.

Videvall, E. *et al.* (2020) 'Host Transcriptional Responses to High- and Low-Virulent Avian Malaria Parasites', *The American Naturalist*, 195(6), pp. 1070–1084. Available at: https://doi.org/10.1086/708530.

Walther, B.A. and Ewald, P.W. (2004) 'Pathogen survival in the external environment and the evolution of virulence', *Biological Reviews*, 79(4), pp. 849–869. Available at: https://doi.org/10.1017/S1464793104006475.

Webb, C.T. *et al.* (2010) 'A structured and dynamic framework to advance traits-based theory and prediction in ecology', *Ecology Letters*, 13(3), pp. 267–283. Available at: https://doi.org/10.1111/j.1461-0248.2010.01444.x.

Wells, K. *et al.* (2018) ‘Global spread of helminth parasites at the human-domestic animal-wildlife interface’, *Global Change Biology*, 24(7), pp. 3254–3265. Available at: https://doi.org/10.1111/gcb.14064.

West, S.A. and Buckling, A. (2003) ‘Cooperation, virulence and siderophore production in bacterial parasites’, *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1510), pp. 37–44. Available at: https://doi.org/10.1098/rspb.2002.2209.

Westerdahl, H. (2007) ‘Passerine MHC: Genetic variation and disease resistance in the wild’, *Journal of Ornithology*, 148(SUPLL. 2), pp. 469–477. Available at: https://doi.org/10.1007/S10336-007-0230-5/FIGURES/2.

Whitfield, J. (2002) ‘Portrait of a serial killer’, *Nature* [Preprint]. Available at: https://doi.org/10.1038/NEWS021001-6.

Wolfson, F. (1936) ‘Synchronicity, periodicity, and the length of the asexual cycle of Plasmodium rouxi in the canary’, *American Journal of Epidemiology*, 23(2), pp. 340–348. Available at: https://doi.org/10.1093/OXFORDJOURNALS.AJE.A118222/2/23-2-340.PDF.GIF.

Woolhouse, M.E.J. and Gowtage-Sequeria, S. (2005) ‘Host Range and Emerging and Reemerging Pathogens’, *Emerging Infectious Diseases*, 11(12), pp. 1842–1847. Available at: https://doi.org/10.3201/eid1112.050997.

Wright, S. (1969) ‘Evolution and the Genetics of Populations. The Theory of Gene Frequencies.’, in *Science*. American Association for the Advancement of Science (AAAS), pp. 722–723. Available at: https://doi.org/10.1126/science.168.3932.722.

8 Acknowledgements

This has been a difficult journey. I have frequently felt isolated, mostly because I isolated myself (though lockdowns certainly didn't help). I have swings of emotions ranging from confidence to anxious incompetence. However, as much as I sometimes avoided it, a meeting with **Olof** always left me encouraged and frequently inspired. During the Geneco mentorship program, I got to hear other students' experiences with their supervisors. I know now that I found a jewel of a supervisor with mine.

Dag, thank you for your insightful answers to my bioinformatic questions, and putting me in touch with the right people.

I've had 3 office mates in succession, and I am grateful to **Anna**, **Aivars** and **Jana** for putting up with my messy desk and all the thinking noises I make. I have greatly enjoyed our philosophical discussions and geography quizzes **Jana.**

A bioinformatician is useless without the work of the experimentalists in the field and lab, so I am indebted to my co-authors and their colleagues and assistants who made this such a rich thesis.

Vincenzo, I have not met another person who has such clarity of thought and expression. Your input has made me a better scientist (I hope). And you're such lovely guy too, which is not a common cooccurrence with brilliance.

Thank you **Kalle** for the last minute Swedish translation of the popular science summary.

So many of my colleagues here on the 2nd floor has greased the wheels of my time here, and I wish I was more open with you all.