A Novel Bioinformatic Approach to Characterise Toll-Like Receptor Gene Diversity in Threatened Birds

---

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Biology at the University of Canterbury

---

Molly Claire Magid

February 2021

Abstract

Species recovery programmes are increasingly using genomic data to measure neutral genetic diversity and calculate metrics like relatedness. While these measures can inform conservation management, determining the mechanisms underlying inbreeding depression requires information about functional genes associated with adaptive or maladaptive traits. Of particular interest is the diversity of toll-like receptor (TLR) genes, which play a crucial role in recognition of pathogens and activation of the immune system. To date, these genes have predominantly been identified and characterised using targeted amplification and sequencing.

In this Proof of Concept, I leverage existing short-read reference genomes, whole-genome resequencing data, and bioinformatic tools to develop a novel method to identify TLR genes and characterise TLR gene diversity. I conduct this Proof of Concept in three stages to characterise TLR gene diversity of three nationally critical birds endemic to Aotearoa New Zealand: tūturuatu/shore plover (Thinornis novaeseelandiae), kākāriki karaka/orange-fronted parakeet (Cyanoramphus malherbi), and kakī/black stilt (Himantopus novaezelandiae).

The highest number of TLRs were identified and characterised in tūturuatu, followed by kākāriki karaka, and then kakī, in which could not characterise any of the TLR genes I identified. Consistent with observations in other threatened species and populations, tūturuatu and kākāriki karaka have relatively low TLR gene diversity in comparison to non-threatened species.

Within the captive tūturuatu population, individuals have a high susceptibility to severe avian pox and a low immune response to vaccination. In contrast, anecdotal evidence suggests the wild tūturuatu population on the Chatham Islands have fewer cases of poxvirus, and when a wild bird does contract the virus, the case is usually mild and the bird is able to overcome infection quickly. The low TLR gene diversity within captive tūturuatu may explain the difficulty captive birds have in overcoming poxvirus infection. These findings will ultimately be used to assess the recent

conservation management action to bring wild eggs from the population on Rangatira to supplement the captive population.

While captive kākāriki karaka are not experiencing a current disease outbreak, the threat of known and emerging pathogens is only growing, and disease may be a concern for the captive population because of the low TLR gene diversity I found. The captive kākāriki karaka population has been supplemented with individuals from wild and translocated populations in an attempt to introduce new, or reintroduce lost, founder lineages. Additional sampling is needed to confirm whether there is TLR gene diversity within the source populations used for supplementation that has yet to be captured, and whether bringing in additional eggs from these populations may contribute novel TLR alleles to the captive population.

I had limited success identifying TLR genes within the kakī reference genome, and no success characterising TLR gene diversity within the semi-wild kakī population. I attribute the former to a relatively incomplete kakī genome, due in part to a relatively high percentage of Ns in the draft assembly. I attribute the latter to missingness within several individuals in the population resequencing data. These combined findings suggest this approach may be less effective for threatened species with relatively incomplete genomes and resequencing data.

The novel approach presented in this thesis provides an example of how functional diversity can be characterised using a bioinformatic approach. As more genomic resources become available for non-model threatened species, as a way to measure genome-wide diversity, these data can be leveraged to characterise functional genes. This has broad implications for how TLR gene diversity can be assessed to inform conservation management actions for threatened bird species in Aotearoa New Zealand and beyond.

Acknowledgements

There are countless people to thank for their work and/or support of this thesis. First, I would like to thank tangata whenua, in particular Hokotehi Moriori Trust, Ngāti Mutunga o Wharekauri, Te Rūnanga o Ngāi Tahu, Te Ngāi Tūāhuriri Rūnanga, Te Rūnanga o Arowhenua, Te Rūnanga o Waihao, and Te Rūnanga o Moeraki for their support.

This research would not have been possible without the conservation practitioners who work with these taonga species, including members of the Kakī Recovery Group (especially Liz Brown and Cody Thyne), the Kākāriki karaka Recovery Group (especially Anne Richardson), and the Tūturuatu Specialist Group (especially Rose Collen, Brett Gartrell, Mirielle Hicks, Dave Houston, Kate McInnes, Leigh Perkasy, Anne Richardson, and Rachel Ward). In particular, I want to acknowledge Brett Gartrell for his guidance and for answering my many questions about all things related to tūturuatu poxvirus infections and immune response.

I extend many thanks to Fulbright New Zealand which funded me to do this research, and to the Fulbright NZ team (especially Magnolia) who supported me in my choice to stay here and always made me feel like I wasn't alone. I'd also like to acknowledge additional financial support from the University of Canterbury.

I could not have completed this thesis without access to the Research Compute Cluster at UC and without support from Francois Bissey. Francois, many thanks for your help troubleshooting technical problems and for your endless generosity in providing more RAM and disk space when I needed it.

Special thanks goes to my supervisor, Tammy Steeves. Thank you so much from taking a chance on me, way back in 2018, when we first talked about this research project. I would never have predicted that it would turn out this way, but I am so happy that it did. When everything with the pandemic was uncertain, you gave me the opportunity to stay and conduct this master's research and your kindness and support meant the world to me. This journey has not always been

straightforward, but through it I have learned so much. This research has taught me much more than just genomics or bioinformatics, it's taught me how to be an engaged scientist who strives for kindness, considers the impacts of my findings, and stands up for what is right, and for that I will always be grateful.

I am so thankful for all the members of the ConSERT team- the support I feel from you all and the laughter and joy we share has been a highlight of this year. Many thanks to Roger Moraga at Tea Break Bioinformatics who helped me from day one with the concept for this thesis and was always willing to solve problems, both big and small. Huge thanks to Ilina for enthusiastically sharing her repository of information about tūturuatu and for doing so with a big dose of humour. Thanks to Levi for helping me puzzle through some of the trickier coding issues I ran into and for providing the tools (like a bright orange hard drive) to run these analyses. I'm grateful to Stephanie for sharing her knowledge of kakī and kākāriki karaka, and for making me feel so welcome when I first arrived. I'd like to thank Nat for sharing with me all of her bioinformatics knowledge and specifics of the kakī genome assembly and for being my first friend out of lockdown. Above all, I extend many thanks to Jana, she has been my mentor through this whole process from helping me write my first line of bash script to editing this thesis, I truly could not have done it without her.

To my parents, thank you so much for your unconditional love and support through this master's. You were the first ones to plant the idea in my head to live and do research in New Zealand. I know that it wasn't always easy for you to let me live so far away during a global pandemic, but I'm beyond grateful that you did. I hope that one day I can share with you the places and the people that I've grown to love here. To Katie, thank you for your kindness and optimism, and thank you for leading the way— after having completed this master's, I am even more in awe of you and how much strength you bring to everything you do. To Sarah, I'm so lucky to have you to share my stories and dreams with, and thank you for always checking in and assuring me things would work out.

To my friends in the US and New Zealand, thanks for your advice and encouragement. To the “tea time revival” group, thanks for challenging me to be a better version of myself and for making me feel like I’m back with you all, drinking tea at our kitchen table. To my “book club” friends, I always leave our meetings with something good— a new perspective, fact, joke, or, more recently, assurances that I can get through anything. Lucy, thanks for alternately commiserating and celebrating with me during these last few months working on our theses, and for helping me to take a break and experience the wildlife and geology of New Zealand. And to the “middle of the road” trivia team, thanks for making sure that I get out and learn new things, even if it means I have to be haunted by the image of a carpet bag steak for the rest of my life.

To Jim Kellner, who started me on my journey of biology research and told me that sometimes the most difficult experiences unexpectedly result in the best that life has to offer, boy do I have a story for you.

Table of Contents

Abstract ii

Acknowledgements iv

Table of Contents vii

List of Figures and Tables ix

Thesis Structure and Chapter Outlines 1

Chapter 1: General Introduction 3

Characterising Functional Genetic Diversity 3

Avian TLR Background 4

Avian Disease and Immunogenetic Diversity 7

Species Description and Natural History 9

Aims of this Thesis 12

Descriptive pedigree statistics for each species 14

Chapter 2: Characterising TLR gene diversity in a captive population of tūturuatu/shore plover vulnerable to poxvirus infection 15

Abstract 15

Introduction 15

Methods 21

Results 31

Discussion 36

Conclusions 40

Chapter 3: Characterising immune gene diversity to inform augmentation of the captive kākāriki karaka/orange-fronted parakeet population 41

Abstract 41

Introduction 41

Methods 45

Results 50

Discussion 54

Conclusions 56

Chapter 4: Challenges associated with an incomplete reference genome and resequencing data prevent characterisation of immune gene diversity in a threatened endemic bird, kakī/black stilt . 58

Abstract 58

Introduction 58

Methods 61

Results 66

Discussion 68

Conclusions 71

Chapter 5: General Discussion and Future Directions 72

Future Directions 74

Data Accessibility 78

References 80

Appendices 89

Contribution Statement 89

Appendix A: Samples Used in Resequencing Analysis 89

Appendix B: Bash scripts associated with TLR identification and characterisation 91

Appendix C: Scripts associated with troubleshooting TLR characterisation in kakī 97

Appendix D: Filtering trials for kakī 100

List of Figures and Tables

Figure 1.1 Visual representation of the conservation management actions informed by this research. Toll-like receptor (TLR) schematics placed above each DNA strand represent potential to increase TLR gene diversity. W= birds of wild origin, T= birds of translocated origin, C = birds of captive origin, SW = birds of semi-wild origin (the vast majority of kakī in the wild are reared in captivity). Bird and egg illustrations provided by Stephanie Galla and Jana Wold, respectively……………………………..…………………….. 13

Table 1.1 Captive tūturuatu and kākāriki karaka pedigree, and semi-wild kakī pedigree, descriptive statistics generated using PMx (tūturuatu: Ilina Cubrinovska unpublished data; kākāriki karaka and kaki: Galla 2019)…… 14

Figure 2.1 Principal component analysis genetic clustering of captive (n=27) and wild (n= 34) tūturuatu. Produced using 50k single nucleotide polymorphisms characterised with reduced-representation data (GBS) using killdeer as a reference genome. It should be noted that the captive outliers within the cluster of wild birds have or are presumed to have wild parentage (Ilina Cubrinovska, unpublished data)………...20

Figure 2.2 Workflow schematic to show how bioinformatic tools and genomic resources are used to identify and characterise TLR genes in threatened birds. See text for details. Abbreviations: TLR= toll-like receptor, ORF=open reading frame, LRR=leucine rich repeat, SNP=single nucleotide polymorphism, bcf=BIM collaboration format. All remaining abbreviations refer to names of bioinformatic tools utilised……………21

Table 2.1 Reference species for toll-like receptor (TLR) identification in tūturuatu. Reference species listed alongside the taxonomic group mostly closely related to tūturuatu. For each reference species, TLR sequences used for comparison were either identified through the use of the NCBI annotation pipeline[1] or through amplification and sequencing of targeted TLRs in the laboratory.………………………………..23

Figure 2.3 SMART protein schematics for each tūturuatu Toll-Like Receptor identified (Available via license: CC BY 2.0). Key shows the visual representation for each protein domain..………………………………...32

Table 2.2 Tūturuatu toll-like receptor sequence specifics and gene diversity statistics for captive tūturuatu population (n=39). Abbreviations: syn:nsyn= ratio of synonymous to non-synonymous SNPs, bp(aa)=base pairs (amino acids)……………………………………………………………………………………………33

Table 2.3 Comparison of tūturuatu TLR gene diversity to species with different threat statuses from the IUCN Red List. The study where these species were analysed is cited at the bottom of the table. Abbreviations: n= number of samples, bp(aa)=base pairs (amino acids), h=number of inferred haplotypes..……………………………………………………………………………………………………....35

Figure 3.1 Pie chart showing wild founder lineages sourced from three mainland valleys represented in captive and translocated populations of kākāriki karaka. The colour of each piece shows the valley that the founder came from and the different shades reflect different founder lineages from that valley. The width of the piece corresponds to the degree of contribution from that founder lineage in the captive population as

of 2018 and in the individuals released on Oruawairua/Blumine during 2011/12. Kākāriki Karaka Recovery Group, unpublished data, reproduced with permission……………………..........................………...43

Table 3.1 Reference species for toll-like receptor (TLR) identification in kākāriki karaka. Species used are listed with taxonomic distance from kākāriki karaka. For each species, TLR sequences used for comparison were either identified through the use of the NCBI annotation pipeline[1] or through sequencing of targeted TLRs in the laboratory…………………..………………………....……………………………………..…...46

Figure 3.2 SMART protein schematics for each kākāriki karaka toll-like receptor identified (Available via license: CC BY 2.0). Key shows the visual representation for each protein domain……………………….50

Table 3.2 Kākāriki karaka toll-like receptor sequence specifics and gene diversity statistics for captive kākāriki karaka population (n=45). Abbreviations: syn:nsyn= ratio of synonymous to non-synonymous SNPs, bp(aa)=base pairs (amino acids)... 51

Table 3.3. Comparison of kākāriki karaka TLR gene diversity to species and populations with different threat statuses. The species are coloured by threat status, and when possible, these statuses come from the IUCN Red List. The study where these species were analysed is cited at the bottom of the table. Abbreviations: n= number of samples, bp(aa)=base pairs (amino acids), h=number of inferred haplotypes, **π** = nucleotide diversity……………………………………………………………………………………….52

Figure 4.1 Founder representation across all breeding individuals over three breeding seasons. Founders with an asterix (*) by their ID have been captured in the captive population over the past three seasons. Reproduced with permission from Galla 2019………………………………………………………………….. 60

Table 4.1 Reference species for toll-like receptor (TLR) identification in tūturuatu, listed alongside the taxonomic group shared with kakī. For each species, TLR sequences used for comparison were either identified through the use of the NCBI annotation pipeline[1] or through sequencing in the lab. Links to the study in which TLRs were sequenced are at the bottom of the table…………………………………………...62

Table 4.2 Summary of bioinformatic tools used for quality tests of the population resequencing alignment for kakī. See text for details. Abbreviations: TLR= toll-like receptor, SNP=single nucleotide polymorphism. All remaining abbreviations refer to names of bioinformatic tools utilised…………………………………...65

Figure 4.2 The first 118 bp of the partial TLR1A sequence shows a 100 bp poly-N repeat…………………66

Figure 4.3 SMART protein schematics for each kakī toll-like receptor identified (Available via license: CC BY 2.0). Key shows the visual representation for each protein domain…………………………………….67

Table A.1 Stud book identification number (SB#) for all tūturuatu, kākāriki karaka, and kakī samples used in this thesis. Parentage of samples is also reported for kākāriki karaka (see Chapter 3 for details). Abbreviations: C=captive, W=wild, T=translocated…………………………………………………..…...89

Table D.1 Filtering trials for kakī vcf file. The filtering step conducted is shown and resulting SNPs in TLR genes are reported. Abbreviations: maf=minor allele frequency, minQ=minimum quality, minDP=minimum depth, maxDP=maximum depth, TLR=toll-like receptor, SNP=single nucleotide polymorphism………...100

Table D.2 Missingness filtering trials for kakī vcf file. Maximum missingness used for filtering is shown (eg. 0.90 = 10% missing data), and resulting SNPs in TLR genes are reported. Abbreviations: maf=minor allele frequency, minQ=minimum quality, minDP=minimum depth, maxDP=maximum depth, TLR=toll-like receptor, SNP=single nucleotide polymorphism…………………………………………………………….... 101

Thesis Structure and Chapter Outlines

Chapter 1: General Introduction

I discuss the existing literature on conservation genomic management, with an emphasis on how existing genomic resources may be utilised to measure functional diversity in threatened species. I also outline the Proof of Concept that I will conduct within this thesis.

Chapter 2: Characterising TLR gene diversity in a captive population of tūturuatu/shore plover vulnerable to poxvirus infection

In this chapter, I describe the novel bioinformatic approach that I use for the Proof of Concept. I use this approach to characterise immune genes in tūturuatu/shore plover, a threatened bird endemic to Aotearoa New Zealand, with a captive population that experiences severe outbreaks of avian pox. I discuss the results in comparison to the TLR gene diversity of other threatened species and to the TLR gene diversity of non-threatened species. I also highlight how these data can be used in future analyses to evaluate whether the conservation management action to augment the captive population with wild-sourced individuals may improve immune gene diversity.

Chapter 3: Characterising immune gene diversity to inform augmentation of the captive kākāriki karaka/orange-fronted parakeet population

In this chapter, I use the same bioinformatic approach to characterise immune gene diversity in kākāriki karaka/orange-fronted parakeet. I discuss the results in comparison to the TLR gene diversity of other threatened species and to the TLR gene diversity of non-threatened species. I also highlight future analyses that can inform whether ongoing supplementation of the captive population leads to increased TLR gene diversity.

Chapter 4: Challenges associated with an incomplete reference genome and resequencing data prevent characterisation of immune gene diversity in a threatened endemic bird, kakī/black stilt

In this chapter, I discuss the challenges that prevented identification of some TLR genes and characterisation of the TLR gene diversity in kakī/black stilt. I describe the methods I used to investigate why these issues may have occurred and suggest that, for kakī, these challenges can likely be attributed to the relative incompleteness of the reference genome and relatively high missingness within the resequencing data.

Chapter 5: General Discussion and Future Directions

Reflecting on the Proof of Concept carried out in this thesis, I describe how future data and genomic technologies can be used to characterise functional gene diversity within other populations of these species and how these future analyses can inform conservation management actions. I also discuss the opportunities and challenges for characterising functional diversity using bioinformatic approaches within bird species of conservation concern.

Chapter 1: General Introduction

Characterising Functional Genetic Diversity

Advances in high-throughput sequencing (HTS) technologies and bioinformatic tools enable scientists to more easily generate and analyse whole genome sequences (Ekblom and Wolf 2014; Auwera et al. 2013). This progress has enabled researchers studying non-model organisms to generate high-quality reference genomes and population genomic datasets, facilitating the use of genomic data to inform the conservation management of threatened species (Primmer 2009; Allendorf, Hohenlohe, and Luikart 2010; Angeloni et al. 2012). For example, genomic data can be used to inform pairing decisions or mating group selection in intensively managed captive or wild populations to minimise mean kinship, and in doing so, minimise inbreeding and maximise founder representation (Galla et al. 2020). Indeed, relatedness metrics based on pedigree, genetic or genomic data are used to guide management decisions under the paradigm of preserving neutral genetic diversity as a proxy for genome-wide diversity (Ivy and Lacy 2012). However, recent studies suggest that neutral genetic diversity may be a poor proxy for functional diversity (van Tienderen et al. 2002; Sommer 2005; Marsden et al. 2013; Grueber et al. 2015). Further, whereas neutral genetic diversity can be used to measure inbreeding and assess inbreeding depression, determining the genetic mechanisms that underlie inbreeding depression requires information about specific functional loci associated with adaptive or maladaptive traits (Kohn et al. 2006; Ouborg et al. 2010; Mable 2019).

To date, in a conservation context, functional genetic diversity has generally been characterized using species-specific primers and PCR protocols to target particular genes (van Tienderen et al. 2002; Evans et al. 2015; Grueber et al. 2015). Then individuals are genotyped and diversity can be estimated at the individual and population levels (Allendorf, Hohenlohe, and Luikart 2010). However, these approaches can be expensive and time-consuming, especially

when a transcriptome is necessary to identify these genes. When whole genome resequencing data and a reference genome for a population of interest already exists, these resources introduce the possibility of using bioinformatic tools to characterise functional genetic diversity (Kohn et al. 2006; O'Connor et al. 2019; Hoelzel, Bruford, and Fleischer 2019). Now is the time to leverage new and existing genomic datasets and explore the potential of genomic resources to characterise functional genetic diversity, particularly in conservation projects with limited resources (Segelbacher and Höglund 2009; McMahon, Teeling, and Höglund 2014).

Bioinformatic tools have been used to identify non-coding RNA or micro-RNA in the genomes of non-human animals (Weber 2005; Copeland et al. 2009). Additionally, there are databases of previously identified functional genes which can guide annotation of these genes within closely-related species (Brucker et al. 2012; Mueller et al. 2020). These and other tools that utilise sequence similarity comparison introduce the possibility that functional genes may be identified within a non-model genome (Grueber 2015; Feng, Stiller, J, and Deng, Y 2020). However only a few studies have used these tools to identify functional genes and characterise functional genetic diversity within a population resequencing dataset (Zhang et al. 2014; Brandies et al. 2020).

Avian TLR Background

Improving immunogenetic diversity within intensively managed species improves outcomes and reduces the burden of wildlife disease (Morris et al. 2015; Zhu et al. 2020). While immunogenetic research on threatened species has mostly focused on MHC genes as an indication of immune health, identifying avian MHC loci and sequences can be technically challenging in non-model organisms (Sommer 2005). MHC genes are extremely polymorphic, highly duplicated, and differ substantially in structure depending on the species (Grueber, Wallis, and Jamieson 2014; Minias et al. 2019). Therefore, there is a growing interest in characterising the diversity of the innate immune toll-like receptor (TLR) genes, which are also involved in

pathogen recognition and immune response (Acevedo-Whitehouse and Cunningham 2006; Vinkler and Albrecht, 2009; Grueber et al. 2012). TLR sequences may be more straightforward to identify using bioinformatic tools since they are relatively conserved and gene duplications are rare and well-defined (Grueber et al. 2015). Studies find that TLR gene diversity tends to be low in small, highly threatened populations compared to large populations of species of least concern (Alcaide and Edwards 2011; Knafler et al. 2017; Dalton et al. 2016; Morrison et al. 2020). Also, TLR gene diversity may not correlate with neutral measures of genome-wide diversity, so directly characterising these genes will reveal how diversity at innate immune genes are affected in highly threatened avian populations (Marsden et al. 2013; Hartmann, Schaefer, and Segelbacher 2014; Grueber et al. 2015).

Toll-like receptors are an ancient part of the innate immune system, present in nearly all multicellular organisms (Singh, Chauhan, and Singhal 2003). They are located either on the cell membrane or intracellularly on the membrane of lysosomes or endosomes in some innate immune or somatic cells (Takeda and Akira 2005). TLRs recognise conserved patterns of pathogens by binding residues, known as antigens, from pathogens that enter the body (Singh, Chauhan, and Singhal 2003). For example, one TLR specific to viruses binds to viral RNA (Medzhitov 2001). Pathogen recognition by TLRs is also necessary for the proper activation and direction of the adaptive immune response (Clark and Kupper 2005; Pasare and Medzhitov 2005; Kawasaki and Kawai 2014).

Avian TLRs were first identified in the chicken (G. gallus domesticus). In non-passerine avian species, there are most often ten avian TLRs, eight of which are orthologous to other vertebrate TLRs (TLR1A/B, TLR2A/B, TLR3, TLR4, TLR5, TLR7), one ortholog to bony fish and Xenopus (TLR21), and one that is unique to reptiles and birds (TLR15) (Alcaide and Edwards 2011; Grueber, Wallis, and Jamieson 2014). Comparison to TLR genes in other animals suggests a pattern of both gene loss and duplication in these regions, like the duplication of vertebrate TLR1 and TLR2 (Kannaki et al. 2010; Temperley et al. 2008). Recent research shows there is a

duplication of TLR7 in some avian taxa. To date, it has been found in Charadriiformes, Cuculiformes, Mesiornithiformes, and some Passeriiformes (Velová et al. 2018). The duplicated TLR7 is thought to have a similar function, though with slight difference, and this is an area of ongoing research (Raven et al. 2017). While some bird species that have a duplication of TLR7 have more than ten TLRs, the overall structure and function of the TLR family remains conserved within the avian phylogeny.

The TLR protein is composed of three protein domains: the extracellular binding domain, transmembrane protein, and intracellular Toll/interleukin 1 (TIR) signaling domain (Yilmaz et al. 2005). There is mostly conserved evolution and synonymous substitutions within TLR sequences, and most variation is witin the region coding for the extracellular binding protein, which is what comes into contact with antigens (Grueber, Wallis, and Jamieson 2014). The binding domain has a conserved pattern of leucine-rich repeats, and so is also known as the leucine-rich repeat (LRR) domain. Inserts of lecuines within the LRR domain may impact pathogen recognition (Offord, Coffey, and Werling 2010), and single nucleotide polymorphisms (SNPs) within this region may also affect the binding affinity of TLRs (Matsushima et al. 2007; Keestra et al. 2008). In contrast, the TIR signaling domain is mostly conserved across the TLR family and is phylogenetically conserved between related species (B. Beutler and Rehli 2002; Yilmaz et al. 2005; Narayanan and Park 2015). The implications are twofold: (1) that the conserved pattern of the LRR domain and conserved TIR domain are ideal for bioinformatic identification due to the similarity of sequences within taxonomically related species, and (2) the sequences within the LRR region are variable and may contain SNPs that are adaptive for species to recognize particular pathogens.

Previous research on avian TLRs have used tools like BLAST (NCBI Basic Local Alignment Search Tool) to design primer sequences to amplify these genes (Chavez-Trevino 2017) and to compare sequences among closely-related species (Raven et al 2017). Further, primers are most often identified through the use of reference TLR sequences from related species (Grueber et al. 2015; Chávez-Treviño et al. 2017; Morrison et al. 2020). In species with

existing whole genome sequences, using established methods for primer design may also help in the identification of the TLR genes (Yilmaz et al. 2005). A recent study on Tasmanian devils has shown the utility of using online genomic databases and comparative genomic tools to identify reproductive genes and characterise functional gene diversity within a population (Brandies et al. 2020). Further, research using genomes from the B10K consortium highlights the efficacy of comparative genomics to identify orthologs and conserved regions using bird species in the same taxonomic class (Feng, S., Stiller, J., and Deng, Y. et al 2020). Increasingly, there are comparative genomics resources available to identify and annotate immune genes within the genomes of non-model species (Grueber 2015; Mueller et al. 2020). These tools provide a path to characterising immune genes using a bioinformatic approach.

Avian Disease and Immunogenetic Diversity

The pathogen pressure on bird species is increasing as they deal with both new and existing pathogens. Birds are increasingly exposed to pathogens due to both habitat loss, which results in more contact with humans and domestic animals, and to more interaction with invasive species (Alley 2002; Daszak 2000; Baron et al. 2014). Consequently, there have been a number of disease outbreaks within threatened bird populations. For example, Psittacine beak and feather disease virus (PBFDV) infected both wild and captive populations of orange-bellied parrots (Raidal and Peters 2018), and here in Aotearoa New Zealand, an outbreak of PBFDV occurred within an isolated subpopulation of red-crowned kākāriki/parakeet (G. J. Knafler, Ortiz-Catedral, et al. 2016). In both cases, the virus was a significant barrier to conservation efforts. Disease may also prove an issue for relocations and translocations of individuals, because an individual brings along the pathogens and microorganisms that they carry into a new population (Ewen et al, 2012). Further, the relocated individual will encounter a new environment with its own community of microbes, and the stress of movement and a new environment may cause immunosuppression and thus greater susceptibility to illness (Alley and Gartrell 2019). Research to determine the

immunological response or disease load within a threatened population only reveals the disease status and immune response of a population retroactively (Alley et al. 2010; Ortiz-Catedral et al. 2011).

Low immune gene diversity increases susceptibility to a variety of pathogens (Spielman et al. 2004). Inbreeding leads to an overall decrease in genetic diversity at immune genes, loss of rare and potentially advantageous alleles, and a decreased ability to adapt to novel or rapidly evolving pathogens (Spielman et al. 2004; Altizer, Harvell, and Friedle 2003). Research from both wild and laboratory populations shows that inbreeding contributes to an increased parasite load, a greater susceptibility to pathogens, a higher likelihood that individuals will act as disease reservoirs, and ultimately higher rates of mortality due to disease (Whiteman et al. 2006; Acevedo-Whitehouse et al. 2003; Ross-Gillespie, O'Riain, and Keller 2007; Whitehorn et al. 2011). One example is the Tasmanian Devil, in which the small, inbred population has very low immunogenetic diversity at genes like TLRs, and in part, this has contributed to an outbreak of transmissible cancer (Morris et al. 2015; Cui, Cheng, and Belov 2015).

Also, increased immune gene diversity may allow individuals to respond quickly and overcome infection (Bonneaud et al. 2012). For example, in Tiēke/South Island saddleback, a parasitic outbreak may have caused selection of one particular variant of TLR1A (Knafler et al. 2017). It is possible that genetic diversity at the TLR1A locus allowed for selection of a variant of TLR1A that leads to a better immune response to parasitic infection. Also, particular TLR alleles have been associated with greater disease resistance, survival, and reproduction (Heng et al. 2011; Davies et al. 2021). Having a robust TLR gene diversity within a population allows for selection of the most beneficial variant to effectively deal with pathogens.

When conservation management programmes for highly-managed threatened species include captive, translocated, and wild populations, it can be possible for captive breeding programmes to be augmented with individuals from diverse source populations, in an effort to minimise the loss of genetic diversity in captivity. Wild or translocated populations, for example,

may retain genetic diversity or represent founder lineages that have been lost from the captive population (Lacy 1987; Doyle et al. 2001; Jiang et al. 2005). Sourcing individuals from genetically diverse or different populations is recognised as a way to improve both genome-wide diversity and immune gene diversity (Grueber et al. 2017; McLennan et al. 2020; Glassock et al. 2021). In this thesis, I conduct a Proof of Concept in three stages to test whether, using a novel bioinformatic approach, I can identify TLR genes and characterise the TLR gene diversity of three threatened bird species in Aotearoa New Zealand. The outcome of this Proof of Concept has broad implications for how TLR gene diversity can be characterised in other bird species to inform conservation management actions.

Species Description and Natural History

Tūturuatu

Tūturuatu (Thinornis novaeseelandiae) is an endangered species in Aotearoa New Zealand. The current wild population is about 240 birds mostly living in the Chatham Islands, and the captive population is about 35 individuals (Department of Conservation Tūturuatu Specialist Group, personal communication). The wild and captive populations have been shown to be genetically distinct from one another (Ilina Cubrinovska unpublished), and the relatively inbred captive population may be suffering from inbreeding depression. In particular, compared to the wild population, the captive population has a higher susceptibility to avian pox (DOC TSG, personal communication). I will characterise the TLR gene diversity within the captive population to determine the baseline TLR gene diversity represented in the conservation breeding programme. This is the essential first step to answer the question: will supplementing the captive population with individuals from the wild lead to an increase in TLR gene diversity?

Photo 1.1: Captive adult tūturuatu. Photo by L. Heyder, reproduced with permission.

Kākāriki karaka

Kākāriki karaka (Cyanoramphus malherbi) is a critically endangered species in Aotearoa New Zealand. The remaining 100-300 wild-breeding birds are restricted to three North Canterbury Valleys (Hawdon, Hurunui and Poulter) and to a translocated population on one predator-free island Oruawairua/Blumine in the Marlborough Sounds (Galla et al 2019). A conservation breeding programme for the species was started in 2003, first with eggs sourced from the Hawdon and Hurunui, and later with birds sourced from the Poulter (Galla 2019). Captive birds were initially released to several predator-free offshore islands and were later released to the mainland, primarily the Hurunui. The current captive population is 107 individuals and is composed of birds with captive parents, and birds brought in as eggs with: (1) wild x wild parentage, (2) translocated x translocated parentage, and (3) captive x wild parentage. The birds with translocated parentage

are sourced from the translocated population on Oruawairua/Blumine. The birds with captive x wild parentage are the offspring of released captive females and wild males in the Poulter Valley. I will characterise the TLR gene diversity of the population in the conservation breeding facility and compare haplotypes of individuals with wild x wild parentage, translocated x translocated parentage, and captive x wild parentage to those with captive x captive parentage. These data will help to answer the question: does sourcing individuals with wild or translocated parentage to augment the captive population lead to an increase in TLR gene diversity?

Photo 1.2: Captive adult kākāriki karaka. Photo by L. Heyder, reproduced with permission.

Kakī

Kakī (Himantopus novaezelandiae) is a critically endangered species in Aotearoa New Zealand with a population of about 169 individuals. The wild population of kakī is mostly restricted to Te Manahuna/Mackenzie Basin and is intensely managed. A conservation breeding and

rearing programme for the species was started in the early 1980s. In this programme, both offspring from captive pairs and eggs harvested from wild pairs are reared in captivity, with the goal of eventually releasing all juveniles into the wild population (Overbeek, 2020; Galla et al. 2020), so for this thesis research the entire kakī population will be referred to as semi-wild. Recent work shows that founder lineages are unequally represented within the semi-wild kakī population, and increasing founder representation may increase genetic diversity (Galla 2019). I will characterise the TLR gene diversity within the semi-wild kakī population to answer the question: does maximising founder representation in the semi-wild population also maximise TLR gene diversity?

Photo 1.3: Captive adult kakī. Photo by L. Brown, reproduced with permission.

Aims of this Thesis

The aims of this research are to use a novel bioinformatic approach with whole genome resequencing data and reference genomes for each threatened bird species to: (1) Identify TLR

genes within the reference genome, and (2) Characterise TLR gene diversity within the captive or semi-wild population of each species described above. Using three species–each with a different natural history and conservation breeding or rearing programme–this research will serve as a Proof of Concept as to how TLR gene diversity, and functional diversity more broadly, can be characterised using a bioinformatic approach. The resulting data will provide an assessment of the current TLR gene diversity, which will inform conservation management actions that may improve TLR gene diversity for the captive populations of tūturuatu and kākāriki karaka and the semi-wild population of kakī (see Figure 1.1).

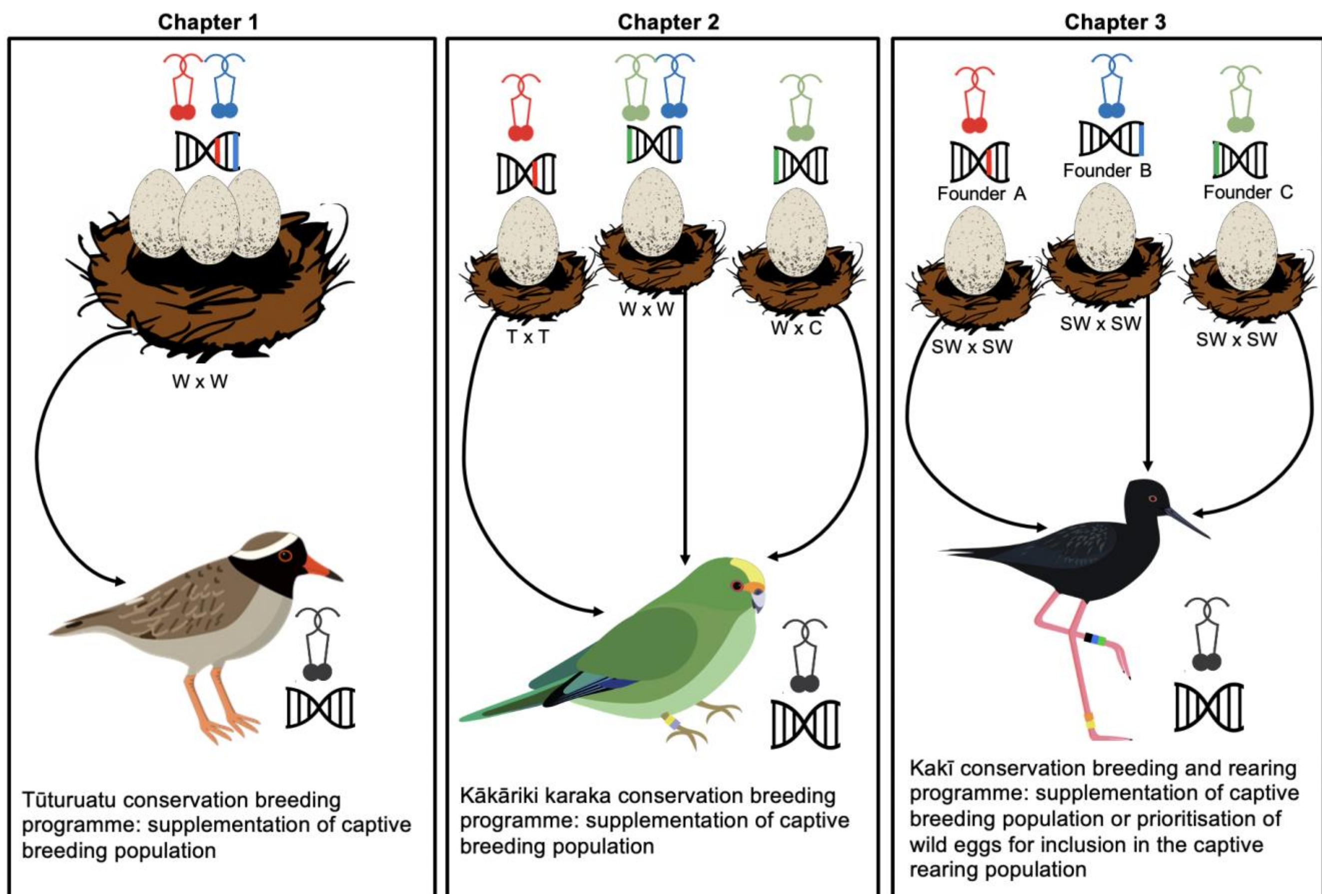

Figure 1.1: Visual representation of the conservation management actions informed by this research. Toll-like receptor (TLR) schematics placed above each DNA strand represent potential to increase TLR gene diversity. W= birds of wild origin, T= birds of translocated origin, C = birds of captive origin, SW = birds of semi-wild origin (the vast majority of kakī in the wild are reared in captivity). Bird and egg illustrations provided by Stephanie Galla and Jana Wold, respectively.

Descriptive pedigree statistics for each species

Table 1.1 lists descriptive statistics for each species based on pedigree information analysed in PMx v. 1.6.20190628 (Lacy, Ballou, and Pollak 2012), which demonstrates similarities and differences between the captive tūturuatu, captive kākāriki karaka, and semi-wild kakī populations. I will refer to this information in each subsequent chapter.

Table 1.1: Captive tūturuatu and kākāriki karaka pedigree, and semi-wild kakī pedigree, descriptive statistics generated using PMx (tūturuatu: Ilina Cubrinovska unpublished data; kākāriki karaka and kaki: Galla 2019)

Diversity Statistic	Tūturuatu	Kākāriki karaka	Kakī
# Individuals in pedigree	1318	624	2680
Sex Ratio (% Males)	0.48	0.50	0.25***
Age Range	<14.3 years	<16.4 years	<24 years
Gene Diversity	0.881	0.92	0.96
# Founders	14*	14	94****
Founder Genome Equivalents	4.2	5.90	12.40
Mean Inbreeding	0.048	0.03	0.03
Average MK	0.119	0.09	0.04
Mean Generation Time	5.32	1.36	4.59
% Ancestry Known	100**	100	55
% Ancestry Certain	100**	100	100
Ne/N	0.513	0.07	0.10

*Two translocated captive-born individuals, subsequently returned to captivity, brought after the original founding of the captive population are counted as founders **Wild founders are counted as having "known" ancestry. ***Sex ratio is skewed as most individuals in the population are unknown. ****Many of these founders were sourced after the captive breeding programme was initiated.

Chapter 2: Characterising TLR gene diversity in a captive population of tūturuatu/shore plover vulnerable to poxvirus infection

Abstract

In the first stage of the Proof of Concept, I leverage existing genomic resources and bioinformatic tools to characterise the TLR gene diversity in the captive population of tūturuatu/shore plover, a nationally critical bird endemic to Aotearoa New Zealand. The captive tūturuatu population has a high susceptibility to avian pox and a low immune response to vaccination. I will present data that shows the captive tūturuatu population has relatively low TLR gene diversity compared to non-threatened birds. The SNP diversity within the TLR genes varies between each kind of TLR, with two TLR genes having the majority of SNP diversity, while the remaining genes have low polymorphism or are monomorphic. Haplotype analysis of the captive population shows that rare alleles are scarce. These data will ultimately be used to determine if supplementing the captive population with individuals from the wild results in an increase in TLR gene diversity and an associated increase in immune response to vaccination.

Introduction

Tūturuatu (Thinornis novaeseelandiae) is a critically endangered species in Aotearoa New Zealand. The current wild population is about 240 birds mostly living in the Chatham Islands, and, at the time of analysis, there were 35 individuals in the captive population (Department of Conservation, Tūturuatu Specialist Group). The conservation breeding programme was established between 1991-1996 with eggs brought from Rangatira Island and was augmented in 2003 by a single adult male from Western Reef (Westy) (DOC TSG).

Since then, the captive population has not been augmented with any additional wild individuals. There was a recent translocation of eggs from Rangatira island to the captive population at Isaac Conservation and Wildlife Trust in December 2020, but these individuals have yet to breed. As a consequence of the small captive population and subsequent loss of founder lineages, the remaining individuals in the captive population are highly related to one another (average mean kinship = 0.119) and so mean inbreeding within the population is relatively high (0.048) (Table 1.1). The population may be suffering from inbreeding depression, because birds in the captive population are highly vulnerable to contracting avian pox.

Avian pox is caused by avipoxvirus (APV), a large dsDNA virus with hundreds of strains that can infect a wide range of bird species (Bolte 1999; Boyle 2007). Some strains of APV are endemic to Aotearoa New Zealand, while others were likely brought in through the poultry industry and by introduced European birds (Ha et al. 2011). An evaluation of both endemic and introduced species in New Zealand revealed that APV is widespread and prevalent in the country (Ha et al. 2011). The common route of transmission is through insect vectors, including sandflies and both native and introduced mosquitoes, which transmit the virus by biting through the bird's skin (King 2003; Ha et al. 2013). A climate-related rise in summer temperatures has allowed for increasing abundance and range expansion of vectors like native and introduced mosquitoes, creating conditions for increased exposure to vector-borne diseases within New Zealand bird populations (Tompkins and Gleeson 2006). APV can also spread through contact with an infected bird, contact with shared objects like feeders used by infected birds, or, rarely, through the inhalation of viral particles (Hansen 1999). Aspects of captive breeding programs like shared aviaries and resources make them particularly vulnerable for the spread of infection (Riper and Forrester 2007).

There are two types of avian pox, cutaneous and diptheric pox infections. Cutaneous pox is the most common infection and causes lesions on the body at the site where it enters the skin and on feather-free areas of the body. Diptheric pox infection is more virulent and causes lesions

within the mouth and respiratory system (Bolte 1999). Here, I focus on the cutaneous version of the infection since, to date, shore plovers have only contracted the cutaneous form of avian pox (Brett Gartrell, personal communication). The pathogenesis of cutaneous avian pox begins when APV enters the skin and causes lesions on the skin. These skin lesions begin to form masses known as pox, which are caused by the rapid proliferation of infected epithelial cells. The cells are hijacked by the virus to replicate so quickly that the center of the mass may be cut off from the blood supply, begin to die, and become necrotic (Brett Gartrell, Personal Communication). Sites of cell death and decay within the pox become prone to fungal or bacterial infection (Riper and Forrester 2007). Secondary infections increase the severity of and rate of mortality from infection (Weli and Tryland 2011; Hansen 1999). Severe cases of avian pox may hinder flight, damage vision, or cause death. Mild pox infection does not cause secondary infections and lasts only a couple of weeks. Regardless of the severity of infection, once individuals recover they gain both humoral and cellular immunity to the virus (Riper and Forrester 2007).

Anecdotal evidence suggests wild tūturuatu contract mild avian pox infections. As compared to individuals in the captive population, the infections appear less severe and shorter and are most often cleared in 1-2 weeks (DOC TSG). In contrast, avian pox infections in captive birds can last for several months and may never resolve (DOC TSG). Longer infection periods allow the pox lesions more time to develop, making it more likely that birds will contract secondary bacterial infections (Brett Gartrell, Personal Communication). While death from the virus remains low, the severity of infection is high. Also, chicks and juveniles are more likely to get sick (DOC TSG), probably because the adaptive immune system is still developing within young birds weeks, especially in terms of acquiring antibody diversity (Fellah, Jaffredo, and Dunon 2008; Palacios et al. 2009). Without a fully developed adaptive immune system, the innate immune system may be especially important for young birds in fighting off poxvirus infections. Protecting juveniles from infection is important for the timely release of young birds into the wild (Dowding 2013, DOC TSG).

Based on the TLR response to poxvirus in other animals, the TLRs that may be most relevant to initiating an innate immune response to APV are TLR3, TLR7, and TLR21 (Brady & Bowie 2014; Cao et al. 2012). TLR3 and TLR7 recognise double-stranded and single stranded viral RNA, respectively, and so they likely detect RNA viral transcripts produced by the APV within a cell during replication. TLR21 recognises dsDNA, and so may detect APV before it begins to replicate. Beyond these specific TLRs, all TLRs are likely involved in recognition of the virus and response to inflammation in some way (Lester and Li 2014).

Notably, poxviruses have strategies to prevent the activation and response of the innate immune system (Giotis et al. 2020). A related poxvirus in humans binds to the TIR signaling domain of a TLR, which prevents that TLR from activating an immune response (Oda, Franklin, and Khan 2011). This may explain how poxviruses can dampen the innate immune response of an affected individual (Brady and Bowie, 2014). TLRs act as the primary alarm for the innate immune system and so may be important to prevent viral replication and establishment (Thompson and Iwasaki 2008). A study of mice infected with smallpox found that early activation of the innate immune system and of inflammatory pathways reduced the severity of the infection, suggesting that a rapid innate immune response to infection may improve disease outcomes (Stanford et al. 2007). Also, activation of the innate immune system is also essential for later initiation of an adaptive immune response, which leads to the development of humoral and cellular memory to pathogens (Pasare and Medzhitov 2005). Since the production of humoral and cellular memory is the goal of vaccination, the innate immune response is also an essential component for a successful vaccination outcome (Pulendran and Ahmed 2006).

In an effort to mitigate avian pox infections, a poxvirus vaccine has been developed to immunise a range of bird species. The first vaccines against avian pox were developed decades ago for use in the poultry industry to protect against fowlpox virus (Romanutti, Keller, and Zanetti 2020). The avipoxvirus vaccine is a live-attenuated virus and is given intradermally at the wing web (D. B. Boyle and Heine 1994). Evaluation of the vaccine in zebra finch proved it is effective

in a passerine and is protective against the two of the strains of APV found in New Zealand. When challenged, vaccinated zebra finches were protected from developing lesions, and post-infection they had a higher antibody count in comparison to birds that were not vaccinated (Ha et al. 2013).

These encouraging results prompted the captive breeding programme to vaccinate captive tūturuatu chicks and juveniles for the past five years. Vaccine trials encompass two rounds of vaccination, the second booster vaccine given two weeks after the first one. This vaccine profile is meant to boost the peak of immunity and memory to infection (Gartrell B, Personal Communication). After the second round of vaccinations, the humoral immunity of each individual is analysed through the use of an ELISA to detect generation of antibodies specific to poxvirus. To date, individuals have not had a strong response to vaccination, comparable to that seen within zebra finch. For most captive birds, there has been little to no antibody production, and for birds that do have some degree of antibody production, it has taken significantly longer to produce this humoral immunity (on the order of months as compared to weeks in zebra finch). If the captive population has a slow or impaired activation of the innate immune system, it would have downstream effects on the adaptive immune system and production of antibodies and memory lymphocytes. An inadequate innate immune response may explain why vaccination efforts have been unsuccessful at producing immunity to APV in tūturuatu.

Previous work based on approximately 50K single nucleotide polymorphisms (SNPs) shows that the wild and captive tūturuatu populations are genetically distinct from one another. (Ilina Cubrinovska unpublished, Figure 2.1). These SNPs reflect neutral, genome-wide diversity, which may also indicate a difference at the gene level, specifically in functional genes like TLRs that contribute to disease resistance. Given that individuals within the wild population have better disease outcomes than captive birds, this may reflect a higher TLR gene diversity within the wild population. If the wild population has a higher TLR gene diversity, then genetic rescue of the captive population through augmentation with wild individuals may improve future disease outcomes and responses to vaccination. However, there is no guarantee that the TLR gene

diversity is higher within the wild population, especially given the significant reduction in the wild population size after mammalian predators were introduced (Dowding and Murphy 2001), so further research is needed to test this hypothesis.

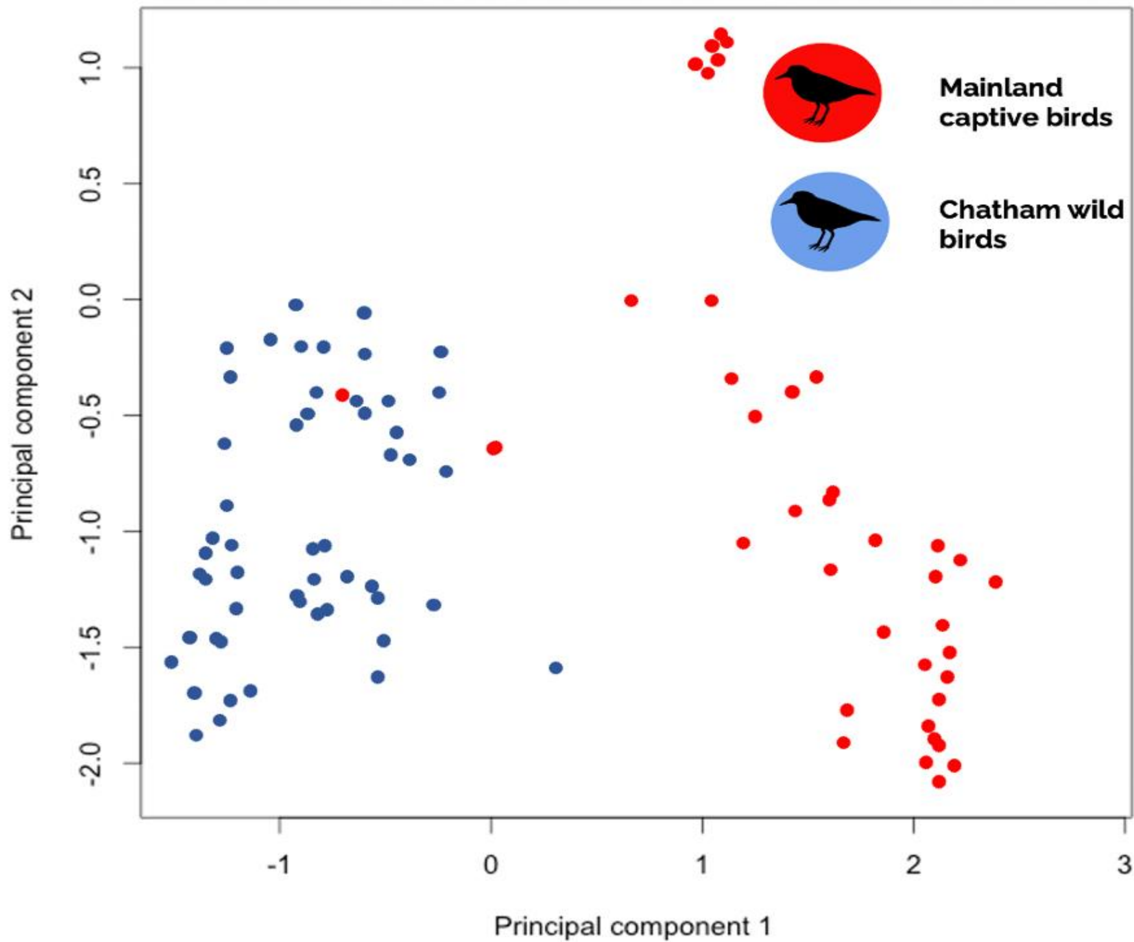

Figure 2.1: Principal component analysis genetic clustering of captive (n=27) and wild (n= 34) tūturuatu. Produced using 50k single nucleotide polymorphisms characterised with reduced-representation data (GBS) using killdeer as a reference genome. It should be noted that the captive outliers within the cluster of wild birds have or are presumed to have wild parentage (Ilina Cubrinovska, unpublished data).

Here, I describe the novel bioinformatic approach used for TLR identification and characterisation and analyse the resulting SNP diversity within the TLR genes of the captive tūturuatu population. This is a first step in evaluating whether the conservation management action of augmenting the captive population with individuals from the wild may result in an increased TLR gene diversity in the captive tūturuatu population.

Methods

The following chart provides an overview of methods used for TLR identification and characterisation (Figure 2.2).

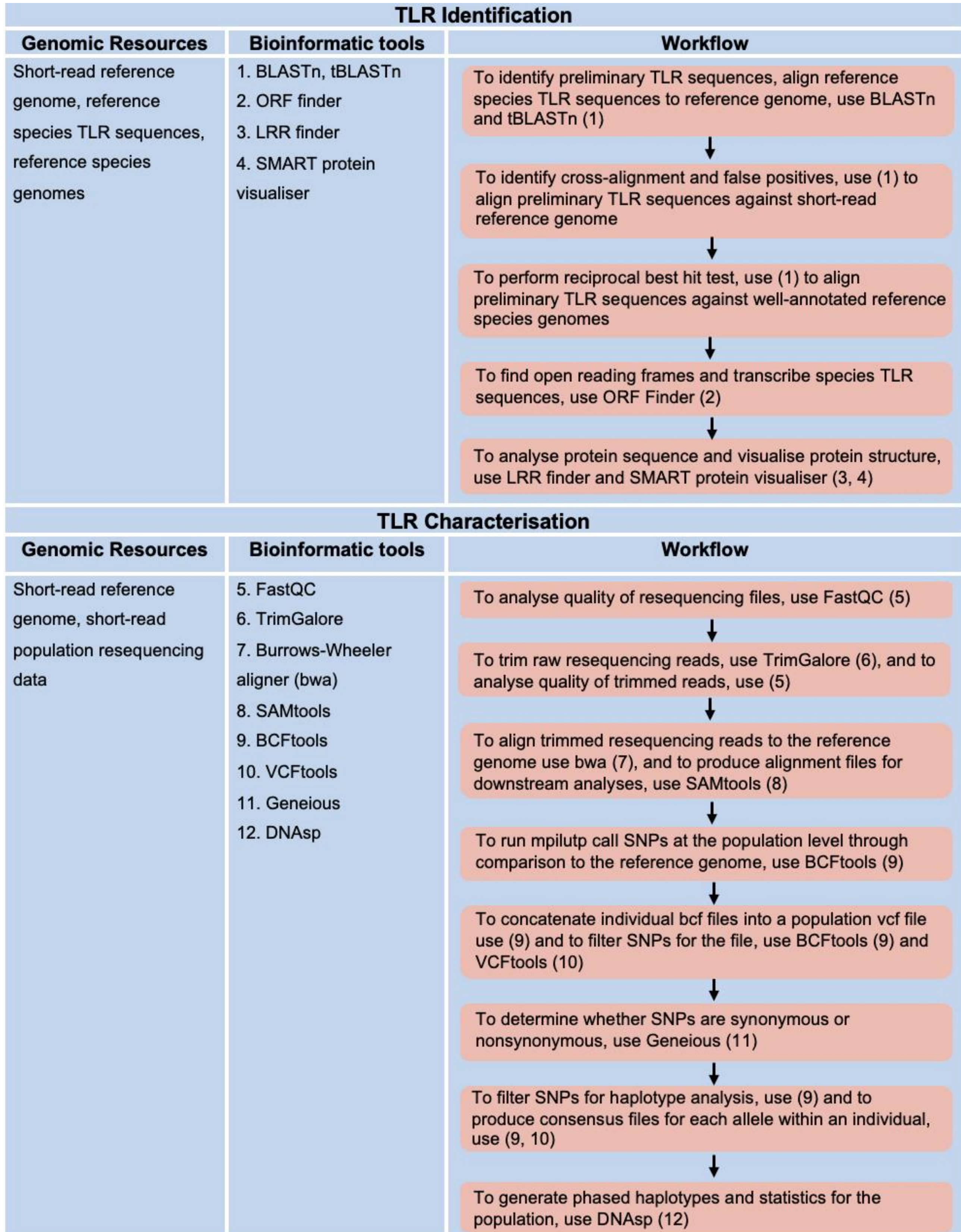

Figure 2.2: Workflow schematic to show how bioinformatic tools and genomic resources are used to identify and characterise TLR genes in threatened birds. See text for details. Abbreviations: TLR= toll-like receptor, ORF=open reading frame, LRR=leucine rich repeat, SNP=single nucleotide polymorphism, bcf=BIM collaboration format. All remaining abbreviations refer to names of bioinformatic tools utilised.

Compute Specifics

For this research, I used a Virtual Machine (VM; 16 vCPUs, 9 TB of memory, 128 GB of RAM) in the host Research Compute Cluster (RCC) at the University of Canterbury.

Reference Genome Assembly

The tūturuatu genome was assembled by Tea Break Bioinformatics (Roger Moraga), and was the genome used to detect SNPs within the captive and wild populations (see Figure 2.1). Initially, the genome was assembled with both Meraculous-2D v. 2.2.5.1 (Goltsman, Ho, and Rokhsar 2017) and the MaSuRCA assembler (Zimin et al. 2013) and these assemblies were compared for quality and completeness. BUSCO v. 3.0.1 (Simão et al. 2015) was used to assess the completeness, duplications, and missingness of the assemblies, and ultimately the MaSuRCA assembly was chosen for the final genome. This genome was assembled with NovaSeq data after the introduction of two-colour chemistry, so there was systematic errors produced within the alignment, namely a high degree of poly-G sequences ("QC Fail Sequencing » Illumina 2 Colour Chemistry Can Overcall High Confidence G Bases" 2016; Moraga, Personal Communication). These errors are difficult to remove because they are detected as high quality sequences and the software made for sequences produced with four-colour chemistry is not primed to detect these irregularities (De-Kayne et al. 2020). The two-colour trimming option was added when trimming raw fastq files, and the filtering pipeline was adjusted to allow filtering for a strand bias adjusted Phred-score. Further changes to the resequencing data were not made at the risk of throwing out SNPs that were accurate, because this analysis is focused on just a few gene regions with low SNP diversity. However, to ensure that the two-colour chemistry errors are fully addressed, the

tūturuatu reference genome is currently being reassembled (Roger Moraga) and characterisation of TLR diversity in tūturuatu will be repeated prior to submission for publication in a peer-reviewed journal. In the meantime, I anticipate impacts to be minimal for the targeted characterisation of TLR genes in tūturuatu.

TLR Identification

BLAST Alignment

Researchers have previously used tools like BLAST (NCBI Basic Local Alignment Search Tool) to design primer sequences that amplify TLR genes (Yilmaz et al. 2005; Chávez-Treviño et al. 2017) and to compare TLR sequences among closely-related species (Raven et al 2017, Mueller et al. 2020). Given the prior uses of BLAST for TLR gene identification, I used BLASTn with known bird TLR nucleotide sequences and tBLASTn with known bird TLR protein sequences, to search for similar nucleotide sequences within the reference genome for tūturuatu (Madden 2013). In this process, I used sequences from bird species with high quality genome assemblies or species which were closely related to tūturuatu. TLR sequences were either identified in the lab through targeted amplification and sequencing or annotated in the reference genome through comparative genomics (Table 2.1).

Table 2.1: Reference species for toll-like receptor (TLR) identification in tūturuatu, listed alongside the taxonomic group shared with tūturuatu. *For each reference species, TLR sequences used for comparison were either identified through the use of the NCBI annotation pipeline[1] or through amplification and sequencing of targeted TLRs in the laboratory.*

Reference TLR Species	**Shared Taxonomic Group**	**Method of TLR Identification**
Killdeer (Charadrius vociferus)	Family Charadriidae	NCBI genome annotation
Black-headed Gull (Chroicocephalus ridibundus)	Order Charadriiformes	Targeted amplification and sequencing[2]

Chicken (Gallus gallus)	Class Aves	Targeted amplification and sequencing[3]
Zebra Finch (Taeniopygia guttata)	Class Aves	NCBI genome annotation excluding targeted amplification and sequencing of TLR4[4]

[1]The NCBI Eukaryotic Genome Annotation Pipeline, https://www.ncbi.nlm.nih.gov/genome/annotation euk/process/ [2]Podlaszczuk et al. 2020, [3]Yilmaz et al. 2005, [4]Vinkler et al. 2009

BLAST uses an algorithm to map input query sequences to the most similar region within the tūturuatu genome. If BLAST is able to map the query sequence to a region or regions within the database, it outputs the region(s) as a list. The list is ordered by the expect value (e-value) of each alignment, which is the number of alignments with a similar score that you'd expect to see by chance. The lower the e-value of the alignment, the less likely that the alignment is due to chance and the more likely it is that the alignement reflects a biological similarity between the sequences (NCBI 2020). In addition to the e-value, the percent query cover and percent identity of the alignment help the user to judge the quality of each alignment. The percent query cover is percent of the query sequence that is aligned to a database sequence, out of the whole length of the query sequence, regardless of the identity of the bases within a sequence. The percent identity is how many of the bases within the query sequence alignment match the bases within the database sequence, out of the whole length of the query sequence. The closer both of these percentages are to 100%, the greater the likelihood that the alignment is biologically relevant.

Relying on BLAST quality information alone would not be sufficient to confirm identification of a TLR gene. I used two further strategies to ensure correct gene identification. The first was that I used multiple bird species to get a more confident call for regions where the reference TLRs were aligning. If the same region in the tūturuatu genome was being mapped to with the same type of TLR sequence from different related species, the greater my confidence was in that call.

Then, I conducted the Reciprocal Best Hit (RBH) test, an approach that has been used within comparative genomics to confirm the identity of orthologs in non-model species using well-annotated reference genomes (RBH) (Kristensen et al. 2011). This limited the test to chicken and

zebra finch, since both have fully sequenced, well-annotated genomes. For the test, I took the "best hit" region identified within tūturuatu using the TLR sequence from a reference species and BLASTed it against the genome of that reference species (Irizarry et al. 2016). If the "best hit" of this BLAST search matches the original reference TLR, it provides greater support that the genes are orthologs of one another (Kristensen et al. 2011). Each preliminary TLR sequence identified in tūturuatu passed the reciprocal best hit test, mapping to the original chicken and zebra finch reference TLR sequences. This helped to confirm not only that I'd likely found a TLR gene, but also helped confirm I had found the specific TLR gene that I was searching for (i.e. if a BLAST search with the preliminary TLR3 sequence brought up TLR3 sequences in reference species).

After conducting these tests of gene alignment, I proceeded with additional quality control steps. I BLASTed each preliminary TLR sequence against the remainder of the tūturuatu genome to find whether there was alignment of this sequence to other regions within the genome. I performed this test to determine whether there were non-target sequences within the tūturuatu genome that may align to these TLR sequences during the process of whole-genome alignment. If there is a high degree of alignment of non-target sequences, then the alignment pipeline should be modified to be more specific and caution may be required in final SNP analysis. I saw little cross-alignment of one type of TLR mapping to other TLR genes (eg. TLR1A reference aligning to the TLR3 region). When this did happen, it was often within pairs of duplicated TLRs (TLR1A & TLR1B, TLR2A & TLR2B), and it was only a partial alignment (at most 50%), so these cross alignments were easy to distinguish from the true alignment. I also looked for false positives within the genome, where the TLR sequences were aligning to other, non-TLR regions. I found no false positive sequences within the tūturuatu genome. Based on these tests, I decided that the Burrows-Wheeler Aligner v0.7.17 (bwa) (Li and Durbin 2009) default maximum mismatch value of 4% for read alignments (allowing only 4% of bases to differ in identity within sequence alignments) would be sufficient to prevent the cross-alignment of duplicated TLRs.

Protein Analysis

Once the preliminary TLR regions passed these quality control measures, I entered each sequence into the NCBI Open Reading Frame (ORF) Finder. This tool searches for reading frames within the query DNA sequence and transcribes the sequence into a protein sequence with each possible reading frame. The result is a graphic showing the protein sequence resulting from each reading frame used transcription (Wheeler et al. 2003). Then, to evaluate protein sequences, I performed BLASTP searches of the entire BLAST protein database with each translated sequence. When I found the correct reading frame, the search would bring up the reference protein TLR sequences that I had originally aligned with.

I used two tools to examine TLR protein sequence and investigate the protein products of the TLR genes. The first tool is LRR finder, which uses a database of toll-like receptor TLR sequences acquired from NCBI to identify the different protein regions within a TLR (LRRs, LRRNT, LRRCT, Transmembrane Protein, TIR Signaling Domain) (Offord, Coffey, and Werling 2010). It uses BLAST-based alignment to compare an input protein sequence to an existing database of TLRs. The tool focuses on predicting potential LRR regions, because this region is the most variable and may affect the binding specificity and affinity of TLRs (Matsushima et al. 2007; Keestra et al. 2008). All searches revealed multiple LRR regions within the sequences, a transmembrane protein region in almost all sequences, and a TIR domain in all sequences. Also the LRR matches within the database always matched the input TLR (i.e. all LRRs for TLR3 were from TLR3 in a reference species).

To further visualise these protein products, I used the tool SMART (Simple Modular Architecture Research Tool). SMART contains a protein database that uses markov modes to identify protein domains within an input protein sequence by calculating the expected value (e-value) SWise score for each alignment between the query sequence and sequences in the SMART database (Schultz et al. 1998; Letunic, Khedkar, and Bork 2021). The protein domains

assigned with low e-values are less likely to be assigned by chance, and at a predetermined e-value threshold, the protein domain will appear as a visual block on a 2D schematic of the protein. This visualisation revealed whether I had captured all expected protein domains within the TLR protein sequences and had not missed parts of the sequence. I was also able to visually compare the tūturuatu TLR proteins to the SMART protein schematics in chicken (Temperley et al. 2008).

Sample Extraction and Resequencing

A total of 39 individuals were resequenced as representatives of the captive tūturuatu population in an aligned project (Ilina Cubrinovska, unpublished data). Brett Gartrell and Isaac Conservation and Wildlife Trust staff collected blood samples during routine health checks. Samples were stored at -80°C until extraction. High quantity and quality DNA was extracted using a tailored lithium chloride extraction method (Galla et al, 2019). Extractions were assessed for quality by running 2 μl of DNA on a 2% agarose gel. A Qubit® 2.0 Fluorometer (Fisher Scientific) was used for DNA quantification.

Libraries for tūturuatu were prepared with the same specifications as used for kākāriki karaka: the Nextera™ DNA Flex Library Prep Kit was used according to the manufacturer's specifications and sequenced across one lane of an Illumina Novaseq™ 6,000 at a coverage depth of approximately 10x (Galla et al, 2020).

TLR Characterisation

Population Resequencing Alignment and Analysis

Once I identified TLR regions within the genome, I performed an alignment of population resequencing data from the captive population to the tūturuatu reference genome to characterise SNP diversity within TLR genes. The goal of alignment is to map population resequencing data to the reference genome and identify all single nucleotide differences between population reads and the reference genome. The alignment is composed of a number of steps that are repeated

for each species, so a custom alignment script was written and modified for use in each species analysed (Appendix A).

First, I ran fastqc on files both prior to trimming to identify low quality regions and baseline quality scores. Then I trimmed raw resequencing data from the captive population using Trim Galore (Kreuger 2020), with clipping for Illumina adapters and --paired and --2colour 20 (the Phred score) as parameters. I added the two-colour chemistry option to prevent incorrect SNP calls downstream, which sets the quality cutoff for all bases except for G bases because G bases are often called as high-quality, even without signal ("QC Fail Sequencing » Illumina 2 Colour Chemistry Can Overcall High Confidence G Bases" 2016). This allowed for trimming and removal of low-quality sequences in non-G bases and for some removal of G bases without a signal, regardless of the quality of those bases (Kreuger 2020). After trimming, I ran fastqc again on the trimmed files to confirm they were of high quality.

Then, I aligned fastq files to the reference fasta file through the use of the Burrows-Wheeler Aligner v0.7.17 (bwa) (Li and Durbin 2009), using a custom pipeline. Initially, I aligned fastq files to a fasta file with only TLR sequences from the reference genome, made using samtools faidx. I used this file in order to limit the alignment of TLR short-read sequences to other regions within the genome and to limit the alignment of non-target short-read sequences to TLR genes in the genome. However, analysis of the final variants showed there was a high degree of missingness in the data, with some individuals missing genotypes at the majority of sites and the population having an average site missingness of 0.372 (on average individuals are missing genotypes at 37.2% of variable sites). After a series of robust filtering steps, there were a very low number of SNPs in the population, and I was concerned that this alignment did not capture all of the genetic variation within these genes. I hypothesized that some sequences within the resequencing data were not aligning to these regions, so I decided to re-align the data to the whole reference genome and compare the resulting alignment files.

The alignment to the whole genome had a higher average depth of coverage within TLR genes, meaning more unique resequencing reads mapped to TLR genes within the whole-genome alignment. The previous alignment had an average depth of ~14x while the alignment to the whole genome had an average depth of ~20x, suggesting there likely was little alignment of TLR reads to other regions in the genome. This difference in read alignment may be due in part to the presence of neighboring sequences within the whole genome that allow for greater mapping of short-read data on the edges of the sequence. Further, there were also few variants with a high depth of coverage (> 200), suggesting there was little alignment of non-TLR short-read sequences to TLR genes. Given its higher quality, whole genome alignment was used for tūturautu and the other two species analysed.

After alignment, SAM files were converted to BAM files and were sorted using SAMtools v. 1.10. (Li et al. 2009). Then a custom perl script ('split_bamfiles_tasks.pl') (Moraga 2018) was used to split the bam files into chunks that could be processed more quickly using BCFtools mpileup (Li et al. 2009). BCFtools mpileup was then run on the chunks of bam files with annotations GT,PL,DP,SP,ADF,ADR,AD, to allow for downstream filtering, which produced bcf files. Then I called variants and produced vcf files using BCFtools. I concatenated the resulting vcf files with BCFtools (Barnett et al. 2011), to examine SNPs on a population level.

I filtered the concatenated vcf file using VCFtools v0.1.16 (Danecek et al. 2011) and determined the parameters for vcf filtering through statistical analysis of a subset of the data. For these analysis I used VCFtools to analyse the alignment for depth per site and per individual, site quality, minor allele frequency (maf), and missingness per site and per individual. Then I visualised statistical output files in R, using ggplot to get an idea of baseline statistics within the SNP dataset. I used these data to set preliminary parameters for filtering and then ran filtering trials with VCFtools v0.1.16 (Danecek et al. 2011) to determine final filtering parameters.

The final parameters I used for filtering are as follows: minor allele frequency (maf) > 0.05, Phred-score (quality) > 20, max-missingness = 0.90, minimum depth > 5, and maximum depth <

200. I also used BCFtools to filter the vcf for strand bias using the parameter strand-bias adjusted Phred-score < 60. Hardy-Weinberg equilibrium filtering was not used since the birds within this study population are in family groups and this violates the assumptions of random mating. I used BCFtools to view the TLR gene regions within the vcf file and used BCFtools stats to examine SNPs within these regions. I analysed final SNPs in Geneious Prime 2020 (https://www.geneious.com/) to determine whether they made synonymous or nonsynonymous changes to the TLR protein products.

Then, for each individual, I generated haplotypes for each TLR gene to determine the contribution of parental alleles, and used these haplotypes to estimate diversity parameters t0 enable a future comparison of TLR gene diversity between the captive and wild populations. To increase quality and decrease errors within the haplotypes, I set sites with low depth and high strand bias to missing using BCFtools +setGT with options -i 'FORMAT/DP<5' and 'FORMAT/SP>60'. Then, I used the filtered vcf file to produce TLR consensus sequences for each individual in the population. I used samtools faidx to target each TLR gene in the genome, and ran BCFtools consensus with parameters -M to output any missing genotypes as "N", and -H 1 and 2 to produce both haplotypes for an individual. If individuals had any sites within the TLR sequence where they were missing a genotype, I did not include them in the haplotype analysis for that TLR gene. While the accepted standard for SNP analysis is that non-synonymous SNPs have significant influence on the functionality of a protein, synonymous SNPs may also change TLR protein function, so all SNPs were used in the construction of haplotypes (Sauna and Kimchi-Sarfaty 2011).

I output resulting TLR consensus sequence for each individual into a population fasta file and repeated the process for every TLR. I imported the resulting fasta files into DnaSP v. 6.12 (Rozas, 2018), and these data were phased for haplotypes using the PHASE algorithm (default settings including recombination). I ground-truthed the resulting haplotypes with an ad hoc assessment of inheritance by comparing parental haplotypes with offspring haplotypes (trios), in

all known families represented in the tūturuatu pedigree. Once I confirmed haplotypes, I analysed them in DNAsp to calculate haplotype diversity, nucleotide diversity (π), and Tajima's D.

Results

There is some uncertainty around the total number of TLRs expected to be found in tūturuatu. Recent evolutionary analysis of the avian TLR family suggests there is a recent duplication of TLR7 within several orders, including Charadriiformes which is the order that tūturuatu belongs to, meaning tūturuatu likely have eleven TLRs total (Velová et al. 2018, Raven et al. 2017).

I identified nine TLRs within the tūturuatu genome: TLR1A, TLR1B, TLR2A, TLR2B, TLR3, TLR4, TLR5, TLR7, and TLR21. I did not identify the whole TLR 15 gene within the genome, but I did identify a partial sequence. Since the sequence ended prematurely at the end of a contig, I believe that the scaffolding of the short-read genome prevented identification of the full gene. I did not find any evidence of the duplication of TLR7 within tūturuatu. After identifying the first gene coding for TLR7, I removed this region from the genome and repeated the BLAST search with the remainder of the genome, to ensure that the reference TLR sequence would not align to the previously identified TLR7 region. However, there was no alignment with any other sequence within the genome. It is possible that the duplication of TLR7 within the reference genome is not fully resolved, meaning the duplicated genes are collapsed into one region. Alternatively, the duplicated region may be incomplete with gaps in the sequence or it may be of low quality, so that it is either partially or fully cut out of the assembly. In the case of a low quality or partial assembly, the duplicated region would not be easily identified through the use of BLAST. In either case, only one TLR7 gene was identified. If the duplication does exist and is of low quality within the tūturuatu genome, this means the population resequencing data may not properly align to the low-quality duplicated TLR7 gene and may instead align to the other TLR7 gene. In this case,

SNPs found within TLR 7 may not be true SNPs, but may instead be artifacts of the duplication reads aligned to this region. Discussion of the data will take this into account when analysing the SNP diversity of TLR7.

The protein schematic structure for each of these TLRs is shown in Figure 2.2. Unlike other analyses that focused solely on the most variable LRR binding region, I decided to analyse the entire sequence for each TLR, because it provides additional confirmation that I had captured the whole TLR and evidence of which TLR I had identified. All TLRs identified display the structure of TLR proteins in other species: a region of leucine rich repeats (the binding domain), then a transmembrane protein (carboxyl-terminal tail) where the TLR sits within the cell or lysosome membrane, and then the TIR (toll-like/interleukin receptor) signaling domain (Yilmaz et al. 2005; Kannaki et al. 2010). While LRRs vary in number between the TLR sequences, the overall structure and approximate amount of LRRs within each type of TLR is similar to what is seen in chicken TLRs (Temperley et al. 2008).

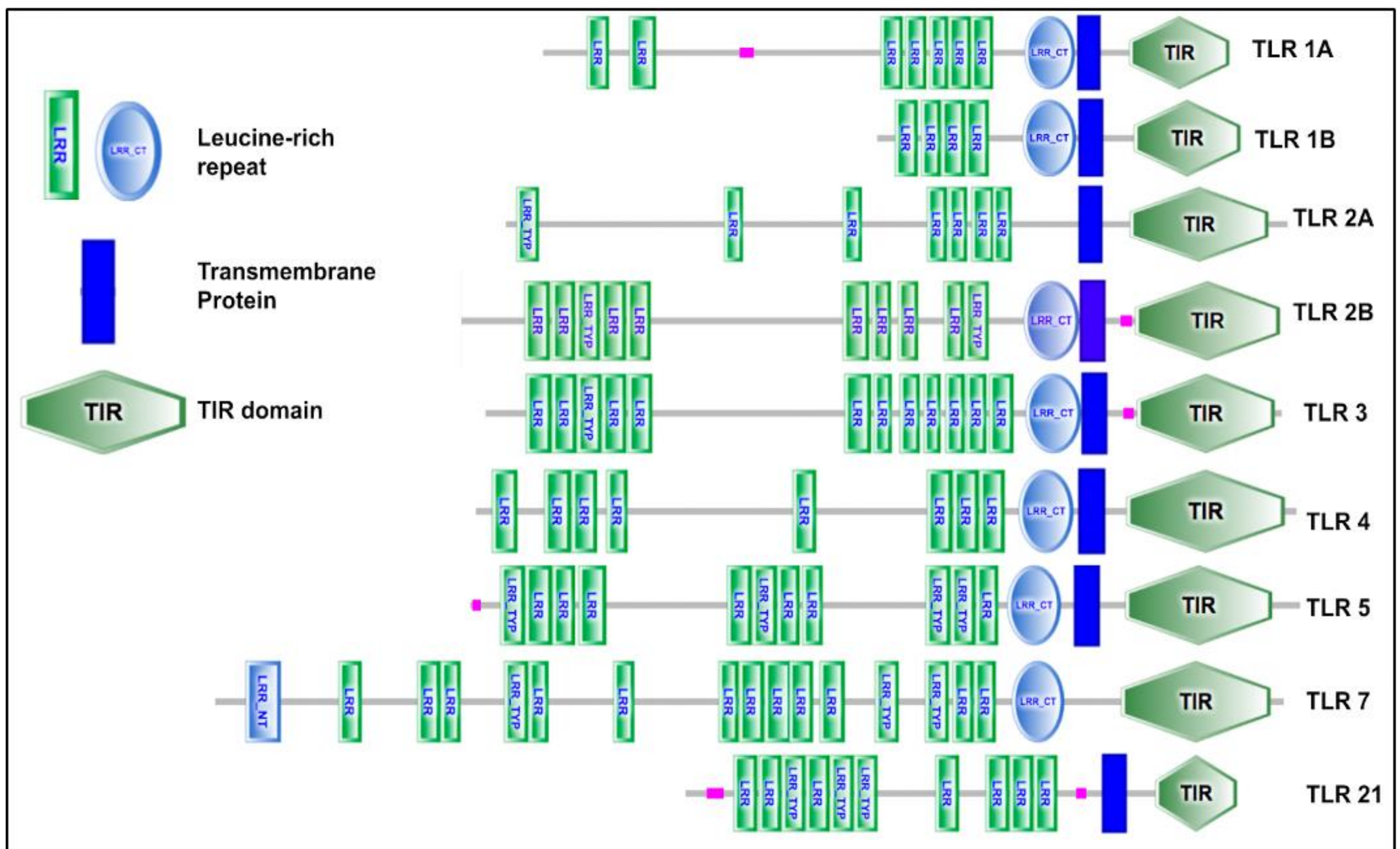

Figure 2.3: SMART protein schematics for each tūturuatu Toll-Like Receptor identified (Available via license: CC BY 2.0). Key shows the visual representation for each protein domain.

There were a total of 29 SNPs in the TLR genes of the captive tūturuatu population (fin=39) (Table 2.2). The SNPs were unevenly distributed among TLR genes, with two that were monomorphic (TLR 1A, TLR 21), two with one SNP each (TLR2A, TLR2B), two with two SNPS each (TLR3, TLR4), one with three SNPs (TLR1A), one with 7 SNPs (TLR5), and one with 13 SNPs (TLR7). Out of the total SNPs, approximately half (15) were non-synonymous, and the remaining (14) were synonymous. The majority of SNPs (25) were within the LRR binding domain and of the SNPs that were in other regions of the TLR protein, three were synonymous and one was non-synonymous.

I observed a total of 25 haplotypes across all TLR genes, with the TLR genes that have higher SNP diversity also having a higher number of inferred haplotypes. Measures of nucleotide diversity were relatively low for most of the loci (mean = 0.000223) and so was haplotype diversity (mean = 0.429). Tajima's D was non-significant for most loci, except for TLR 5 and TLR 7, where the value was significant ($p < 0.05$) and positive for both.

Table 2.2: Tūturuatu toll-like receptor sequence specifics and gene diversity statistics for captive tūturuatu population (n=39). Abbreviations: syn:nsyn= ratio of synonymous to non-synonymous SNPs, bp(aa)=base pairs (amino acids).

Toll-like receptor	Specificity[1]	SNPs (syn:nsyn)	Length bp (aa)	# Birds haplotyped	# Inferred haplotypes	Nucleotide diversity (π)	Haplotype diversity	Tajima's D
TLR 1A	Bacteria, liproproteins	3(1:2)	2190 (728)	36	4	0.00033	0.457	0.35449
TLR 1B	Bacteria, liproproteins	0	1087 (361)	-	1	-	-	-
TLR 2A	Bacteria, triacylated liproproteins	1(1:0)	2423 (806)	38	2	0.00018	0.042	1.33218
TLR 2B	Bacteria, triacylated liproproteins	1(1:0)	2156 (717)	39	2	0.00017	0.360	0.93281
TLR 3	dsRNA	2(2:0)	2237 (746)	37	3	0.00026	0.467	0.31418

TLR 4	LPS, gram negative bacteria	2(1:1)	2259 (752)	36	3	0.00024	0.494	0.58182
TLR 5	Flagellin	7(4:3)	2585 (860)	34	5	0.00123	0.548	3.01252**
TLR 7	ssRNA	13(4:9)	3141 (1045)	39	4	0.00026	0.639	4.05276***
TLR 21	Microbial DNA	0	1649 (548)	-	1	-	-	-

[1]Alcaide and Edwards 2011, *p<0.05, **p < 0.01, ***p < 0.001

It is difficult to directly compare TLR gene diversity between different bird species due to a number of factors including differences in natural history, captive/translocated/wild origin, pathogen exposure, sample size, and population size. Despite these issues, comparing TLR diversity between species with different threat statuses allows for some degree of inference about how threatened species compare to common species in terms of TLR gene diversity (Dalton et al. 2016; Morrison et al. 2020). Here, I report only the SNPs within the LRR binding domain for tūturuatu, because this was the only TLR region analysed in species used for comparison. I found SNPs within the LRR binding domain by analysing the entire LRR region within the protein visual for each TLR (see Figure 2.2), but because I did not use tailored TLR primers for the LRR binding domain (as comparison studies did), these regions may be slightly different.

The species chosen for comparison to tūturuatu were selected from the small number of potential birds in which TLR gene diversity has been analysed. I included species with different IUCN Red List statuses to best examine how having a small, threatened population may influence TLR gene diversity compared to species of least concern. The priority was to identify species that were closely related to tūturuatu, so the black-headed gull was chosen as a species of least concern, but because not all TLRs were identified in this species, the lesser kestrel was also included as a species of least concern. Unfortunately there were no closely-related threatened species in which TLRs have been studied, so the mohua were chosen as a species endemic to Aotearoa New Zealand. The TLR dataset for mohua was also incomplete, so the Stewart Island robin was chosen as another threatened species for comparison. While the Stewart Island robin is listed as a species of least concern, the population is isolated and had experienced two

population bottlenecks at the time of analysis, and so more accurately represents a threatened population within Aotearoa New Zealand (Grueber et al. 2012). The resulting comparison is shown in Table 2.3.

Table 2.3: Comparison of tūturuatu TLR gene diversity to species with different threat statuses from the IUCN Red List. The study where these species were analysed is cited at the bottom of the table. Abbreviations: n= number of samples, bp(aa)=base pairs (amino acids), h=number of inferred haplotypes.

TLR Locus	Species	n	Length bp (aa)*	Threat Status	# SNPs	h
TLR 1A	Tūturuatu (Thinornis novaeseelandiae)	39	1691	Endangered	3	3
	Mohua (Mohoua ochrocephala)[1]	21	987	Endangered	3	4
	Stewart Island Robin (Petroica australis rakiura)[2]	10	1166	Least Concern*	2	2
	Lesser Kestrel (Falco naumanni)[3]	8	1163	Least Concern	19	11
TLR 1B	Tūturuatu (Thinornis novaeseelandiae)	39	706	Endangered	0	1
	Mohua (Mohoua ochrocephala)	21	897	Endangered	3	4
	Stewart Island Robin (Petroica australis rakiura)	10	971	Least Concern*	3	2
	Black-headed gull (Chroicocephalus ridibundus)[4]	60	1044	Least Concern	19	26
	Lesser Kestrel (Falco naumanni)	8	990	Least Concern	16	15
TLR 2A	Tūturuatu (Thinornis novaeseelandiae)	39	1779	Endangered	1	2
	Stewart Island Robin (Petroica australis rakiura)	10	1034	Least Concern*	1	2
	Lesser Kestrel (Falco naumanni)	8	543	Least Concern	6	5
TLR 2B	Tūturuatu (Thinornis novaeseelandiae)	39	1554	Endangered	1	2
	Stewart Island Robin (Petroica australis rakiura)	10	1021	Least Concern*	5	3
	Lesser Kestrel (Falco naumanni)	8	565	Least Concern	2	3
TLR 3	Tūturuatu (Thinornis novaeseelandiae)	39	1408	Endangered	2	3
	Mohua (Mohoua ochrocephala)	23	920	Endangered	1	2
	Stewart Island Robin (Petroica australis rakiura)	9	1087	Least Concern*	0	1
	Black-headed gull (Chroicocephalus ridibundus)	60	1128	Least Concern	11	16
	Lesser Kestrel (Falco naumanni)	8	1160	Least Concern	1	2
TLR 4	Tūturuatu (Thinornis novaeseelandiae)	39	1660	Endangered	1	2
	Mohua (Mohoua ochrocephala)	24	660	Endangered	2	3
	Stewart Island Robin (Petroica australis rakiura)	10	649	Least Concern*	4	5
	Black-headed gull (Chroicocephalus ridibundus)	60	831	Least Concern	8	7
	Lesser Kestrel (Falco naumanni)	8	818	Least Concern	6	7

TLR 5	Tūturuatu (Thinornis novaeseelandiae)	39	1885	Endangered	5	3
	Mohua (Mohoua ochrocephala)	20	1035	Endangered	11	6
	Stewart Island Robin (Petroica australis rakiura)	10	1229	Least Concern*	2	3
	Black-headed Gull (Chroicocephalus ridibundus)	60	1203	Least Concern	14	38
	Lesser Kestrel (Falco naumanni)	8	1265	Least Concern	20	16
TLR 7	Tūturuatu (Thinornis novaeseelandiae)	39	3141	Endangered	12	3
	Stewart Island Robin (Petroica australis rakiura)	10	1010	Least Concern*	3	2
	Black-headed Gull (Chroicocephalus ridibundus)	60	762	Least Concern	16	49
	Lesser Kestrel (Falco naumanni)	8	869	Least Concern	3	4
TLR 21	Tūturuatu (Thinornis novaeseelandiae)	39	1649	Threatened	0	1
	Mohua (Mohoua ochrocephala)	20	641	Threatened	0	1
	Stewart Island Robin (Petroica australis rakiura)	10	618	Least Concern*	3	4
	Lesser Kestrel (Falco naumanni)	8	834	Least Concern	12	3
Total SNPs (syn:nsyn)	Tūturuatu 25 (11:14)	Mohua 20 (11:9)**	Stewart Island Robin 23 (9:14)	Lesser Kestrel 85 (53:32)	Black-headed Gull 68([a])**	
Total haplotypes	20	20**	24	66	136**	

References: [1]Grueber et al 2015, [2]Grueber et al 2012, [3]Alcaide and Edwards 2011, [4]Podlaszczuk et al. 2020 *Listed as a species of least concern, but the population in this study is small and isolated (Grueber et al. 2012.) **Data are incomplete for this species, so the total reflects 5 TLRs [a]syn:nsyn ratio is unknown.

As compared to species of least concern, tūturuatu had much fewer total SNPs and had a similar amount of SNPs as the Stewart Island robin and mohua. The ratio of synonymous to non-synonymous SNPS was lower in tūturuatu than in both species of least concern. Tūturuatu had a similar amount of SNPs and a similar ratio of synonymous to non-synonymous SNPs as the other threatened species did. Tūturautu also had a similar number of total haplotypes within TLR genes as mohua and the Stewart Island robin did. Total haplotypes are much higher for the species of least concern as compared to threatened species.

Discussion

The first stage of this Proof of Concept was realised, because I was able to use existing genomic resources and a bioinformatic approach to identify and characterise TLR sequences in

captive tūturuatu. These results affirmed the underlying premise of this thesis research and demonstrated the utility of the bioinformatic approach that I developed. Subsequently, I applied the approach, outlined in this chapter, in the remaining two stages of this Proof of Concept (Chapter 3, Chapter 4).

I found that the TLR gene diversity within tūturuatu is low in comparison to species of least concern and is consistent with what is seen in other threatened species with small populations (Grueber et al. 2015; Dalton et al. 2016; Morrison et al. 2020). There were about the same number of non-synonymous as synonymous SNPs in tūturuatu, which was also found in a number of other threatened endemics in New Zealand (Grueber et al. 2015). There were fewer haplotypes within tūturuatu than in the species of least concern, which reflects the low SNP diversity in TLR genes. Nucleotide diversity and haplotype diversity were also relatively low for most TLR genes. The value of Tajima's D for both TLR5 and TLR7 were significant and positive, indicating that rare alleles are scarce. The positive significant values indicate a rapid decrease in population size and/or balancing selection (Tajima 1989). Of the SNP diversity that exists, there is an uneven distribution of this diversity within TLR genes. In particular, two TLRs (TLR5 and TLR7) have a higher SNP diversity than the other TLR genes, some of which are monomorphic.

Here, I provide a preliminary assessment of the TLR SNP diversity found within tūturuatu and how specific TLRs may be important for response to poxvirus infection and vaccination. However, I acknowledge that further analyses will be necessary to adequately test these hypotheses.

TLR5 is specific for flagelin, meaning it most often recognises pathogenic bacteria that enter the body (Feuillet et al. 2006). While TLR5 may not directly recognise poxvirus, it may still have a role in responding to inflammation and activation cues (Brady and Bowie 2014). Additionally, it is possible that SNP diversity within TLR5 may be important for response to secondary bacterial infections observed within some individuals with avian pox (Hendaus, Jomha, and Alhammadi 2015).

The SNP diversity within TLR7 was the highest of all TLR genes. However, it is important to note that this result may not reflect the true diversity of TLR7 within the population, if there is a TLR7 duplication that is not well-resolved within the reference genome. In this case, there may be a lower SNP diversity within TLR7 than found here. For this analysis, I will interpret these results with caution, but future analyses should investigate whether there is a duplicated TLR7 gene in tūturuatu.

TLR7 is specific to viral RNA and may be important in the innate immune response to poxvirus. While APV is a DNA virus, once it begins the process of replication through the use of a viral RNA polymerase, it produces RNA transcripts of the virus within the cell (Brady and Bowie 2014). TLR7 is located on the endosome membrane and could bind to RNA transcripts resulting from replication, which would alert the immune system to respond. Another viral-sensing TLR that recognises dsRNA is TLR3, and in the captive population, TLR3 has a small amount of SNP diversity and may also be important in recognising these viral transcripts. However, even though TLR3 and TLR7 may trigger a response from the innate immune system, the activation of the immune system may be too late, because at this point the APV has already started replicating (Giotis et al. 2020). Once replication begins, the virus is able to proliferate widely and travel through the bloodstream to other parts of the body, and this widespread infection is more difficult for the immune system to fight.

The TLR that may detect APV early on is TLR 21, because it is specific for microbial DNA. A study in mice found that wild-type mice infected with poxvirus had much higher survival rates than TLR9-deficient mice (TLR 9 is the mammalian equivalent to TLR21) (Samuelsson et al. 2008). Further, TLR9-deficient mice were more likely to have symptoms of a pox infection in comparison to some wild-type mice who were able to clear the virus before the onset of symptoms (Samuelsson et al. 2008). This suggests that TLR21 may play an important role in early detection of APV and early response of the innate immune system, which may be important for both a decreased susceptibility to infection and for greater survival. In captive tūturuatu, there was no

SNP diversity within the TLR21 gene. This means there is only one allele at TLR21 in the whole captive population and this allele may not properly recognise APV or activate the immune system. Without early detection of the virus, an infected bird may have a longer and more severe pathogenesis. The theory that diversity at one specific TLR may be important, is consistent with a study that found diversity at specific TLR genes was more important than overall SNP diversity within TLR genes (Bateson et al). This means that while there is some TLR SNP diversity overall, it may not be within the genes that are most important for fighting off infection.

Further, APV has a number of strategies to evade and slow down the innate immune system. Specifically, it targets the signaling pathways of all TLRs in the innate immune system by producing agonists A52 and A46. A52 is an inhibitor of NFkB, a transcription factor that increases the expression of pro-inflammatory genes, like those that release cytokines (Liu et al. 2017). A46 directly inhibits TLR signaling by interfering with the TIR signaling domain and with the adaptor molecules that are necessary for signal transduction and downstream innate immune processes (Brady and Bowie 2014; Harte et al. 2003; Stack et al. 2005). The TIR domain is conserved across all types of TLRs, which means that the virus can decrease immune function across the board, which may explain why secondary infections are prevalent among infected birds. Also, the majority of SNPs are within the binding region of the TLRs, so it is unlikely that there are alleles variation in TIR which could evade these viral agonists. A slow or impaired innate immune response may contribute to a delayed or impaired adaptive immune response (Cao et al. 2012).

Without an adequate adaptive immune response, individuals are unable to develop antibodies or memory lymphocytes to pathogens. This may explain the lack of humoral response to vaccination seen within captive tūturuatu. Results from a study of smallpox vaccination showed that low responders, those who failed to develop sufficient humoral and cellular immunity, also had a lower activation of innate antiviral immune genes and a lower release inflammatory molecules, which in part activate the adaptive immune system (Kennedy et al. 2016). These results indicate that the innate immune system may play a role in the absence of response to

vaccination. Even when an individual is immunocompetent, APV has multiple ways to handicap both an innate and adaptive immune response, so it is likely there are multiple reasons behind the lack of response to vaccination (Seet et al. 2003; Johnston and McFadden 2003). Further investigation into whether and/or how captive tūturuatu are mounting a response to poxvirus infection is required to determine which immune cells are responding to poxvirus and which genes may be the most important for fighting avian pox infection.

Conclusions

The novel bioinformatics approach presented in this chapter was successful for identifying TLR genes and characterising TLR gene diversity in tūturuatu. Relative to non-threatened species, results from the captive tūturuatu population show that overall SNP diversity within TLR genes is low. Higher SNP diversity in TLR 7 may be important for an immune response to poxvirus, but the low diversity in TLR 21 may mean that the immune system is delayed in responding to infection, leading to a longer and more severe pathogenesis. The current efforts to prevent disease like vaccination are not sufficient to protect against infection, so there is a need to focus on improving immune function in the captive population. Future research, also applying this approach, will be used to characterise TLR gene diversity in the wild population and determine whether supplementation with individuals from the wild will improve TLR gene diversity and immune function in the captive population.

Chapter 3: Characterising immune gene diversity to inform augmentation of the captive kākāriki karaka/orange-fronted parakeet population

Abstract

In the second stage of this Proof of Concept, I apply the same bioinformatic approach to determine whether I can characterise TLR gene diversity in kākāriki karaka/orange-fronted parakeet. Kākāriki karaka is a threatened bird species endemic to Aotearoa New Zealand that is distantly related to tūturuatu, but unlike tūturuatu the kākāriki karaka captive population has been periodically augmented from populations outside captivity since it was established. The approach to characterise TLR gene diversity is not as straightforward for kākāriki karaka because of three low-depth and high missingness individuals that I ultimately remove from analysis. Despite these issues, I will present data that show the TLR gene diversity within the population is low and similar to that within other threatened species and populations. I am not able to accurately identify haplotypes and haplotype diversity, and therefore cannot determine whether bringing in eggs with wild x wild, translocated x translocated, or captive x wild parentage contributes novel haplotypes to the captive population. Further analysis of the captive population as well as of all source populations for augmentation will help determine whether all existing TLR gene diversity within the species is being captured in the captive population.

Introduction

Kākāriki karaka/Orange-Fronted Parakeet (Cyanoramphus malherbi) is a critically endangered endemic species in Aotearoa New Zealand. In te reo Māori, kākāriki karaka translates as “little orange parrot” (https://maoridictionary.co.nz/), referring to the orange strip on

top of the beak that distinguishes them from the red-crowned and yellow-crowned kākāriki. While previous research suggested that kākāriki karaka may be a hybrid of the yellow-crowned kākāriki, molecular and genetic analyses confirm that this species is distinct from the other New Zealand kākāriki (Kearvell et al 2002; Kearvell and Grant 2003; Rawlence et al. 2015).

There is little written history of the species, but it is thought to have been widely distributed, according to reports from the 1800s (Kearvell et al 2003). Their range decreased dramatically during the 19th and 20th centuries because of threats from introduced mammalian predators like rats and stoats and from habitat loss (Kearvell and Legault 2017). Kākāriki karaka are now the rarest forest bird and the rarest parakeet in New Zealand with a wild population of only 100-300 individuals (Department of Conservation, https://www.doc.govt.nz). In the wild, birds are currently restricted to beech forests in three North Canterbury Valleys (the Hawdon, Hurunui and Poulter) and one predator-free island (Oruawairua/Blumine) in the Marlborough Sounds (Department of Conservation, Recovery Group). A conservation breeding programme for the species was started in 2003, after two years of rat plagues in 2001 and 2002 caused dramatic population declines (Bird Life International 2018). The captive population was started with fourteen eggs harvested from the Hawdon and Hurunui. The programme initially released individuals on several predator-free offshore islands, but subsequently most offspring are released to mainland sites, mostly in the Hurunui (DOC RG).

The only remaining translocated population is on Oruawairua/Blumine which was founded in 2011/12 with 62 offspring from captive pairs (DOC RG). A retrospective evaluation of the 2011/12 translocations showed that, compared to the 2018 captive population, 10 of the 12 founder lineages represented in the 2018 captive population were represented in the translocated population. The ten shared founder lineages are unequally represented in both the captive and wild populations, and the degree to which each lineage is represented within the translocated population is different from the degree to which each lineage is represented within the captive population (Figure 3.1, DOC RG unpublished). This suggests that there may be alleles retained

within the translocated population that have been lost from the captive population and vice versa (Galla and Steeves 2018).

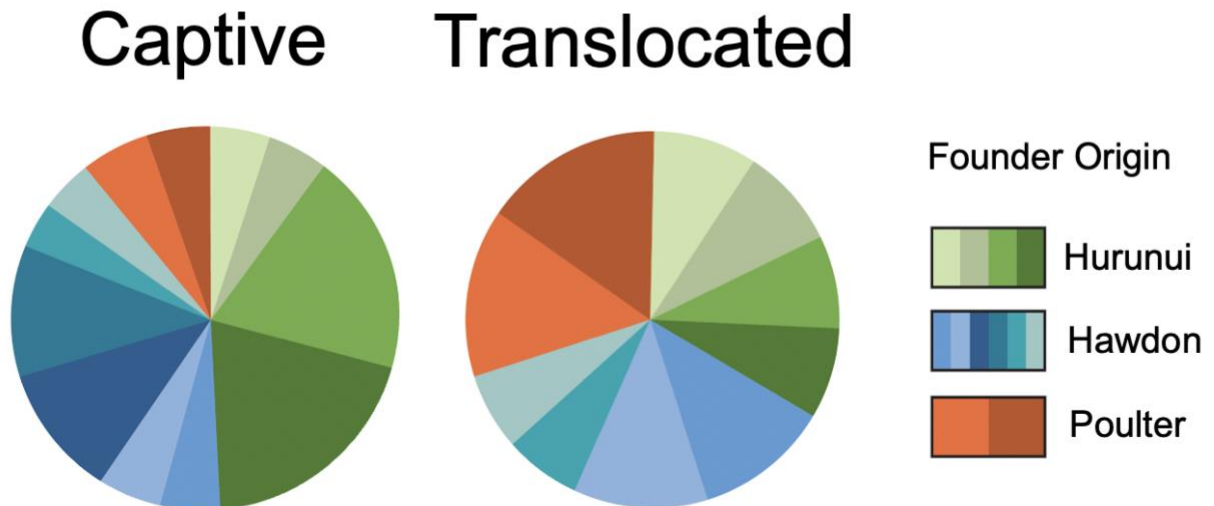

Figure 3.1: Pie chart showing wild founder lineages sourced from three mainland valleys represented in captive and translocated populations of kākāriki karaka. The colour of each piece shows the valley that the founder came from and the different shades reflect different founder lineages from that valley. The width of the piece corresponds to the degree of contribution from that founder lineage in the captive population as of 2018 and in the individuals released on Oruawairua/Blumine during 2011/12. Kākāriki Karaka Recovery Group, unpublished data, reproduced with permission.

A focus of the captive breeding programme for kākāriki karaka is to capture all of the existing genetic diversity within the wild and translocated populations. To augment the captive population, eggs were sourced from wild pairs in the Poulter and the Hawdon and from translocated pairs on Oruawairua/Blumine (Galla et al. 2020). There was also a one-time release of captive females in the Poulter, so that they could pair with lone wild males, and the resulting eggs were brought into captivity. Therefore, the current captive population is composed of birds with two captive parents, and birds brought in as eggs with: (1) wild x wild parentage, (2) translocated x translocated parentage, and (3) captive x wild parentage. While eggs were brought

from outside captivity to increase neutral genetic diversity, it is unknown whether TLR gene diversity has also increased as a result of these conservation management actions.

While currently there are no known outbreaks in captive or wild kākāriki karaka populations, there are potential disease threats to kākāriki karaka within Aotearoa New Zealand. For example, Psittacine Beak and Feather Disease (PBFD) can infect both Old World and New World parrots, and is thought to have been introduced to New Zealand from the import of exotic parrots from Australia like the Eastern Rosella (Mander et al, 2003). Beak and feather disease virus (BFDV) has been shown to infect the congeneric red-crowned kākāriki, though the disease appears to be mild (Ha et al. 2009). A study of a red-crowned kākāriki population on Little Barrier Island indicates that TLR gene diversity may be important for adaptation to PBFD, since the prevalence of BFDV decreased over time and there appeared to be selection within TLR3 that may have been related to the disease (Knafler et al. 2016). To date, there has not been a documented case of PBFD within kākāriki karaka (Massaro et al. 2012; DOC RG), so it is unclear how the virus may affect them. Other common diseases like Aspergillosis are known to affect kākāriki karaka (DOC RG), and there are pathogens like psittacine pox brought by invasive species which may spill over into native Psittacines (Jackson, Morris, and Boardman 2000).

Conservation breeding programmes are at an increased risk of disease outbreaks because birds are often kept in close quarters and resources often shared. While these programmes are able to deal with the risk of disease through implementation of quarantines and deep cleaning regimes, infections within captive populations may still persist and prevent translocations of captive birds to the wild (Ballou 1993; Ewen et al 2012). Further, the global risk of emerging wildlife disease is increasing, driven by invasive species and increased exposure to domestic animals or other wild populations (Tompkins et al. 2015). In addition to current disease prevention protocols, conservation breeding programmes can begin to actively consider improving immunogenetic diversity within captive populations to help them respond to both common and novel pathogens (Glassock et al. 2021). Periodic augmentation of the kākāriki

karaka conservation breeding programme is intended to capture wild lineages not well-represented (Poulter), or to recover founder lineages that may have been lost(Oruawairua/Blumine), within the captive population (Galla 2019). However, it is unknown how these management actions may affect the TLR gene diversity within the captive population.

In this chapter, I apply the same bioinformatic approach used in Chapter 2 to characterise TLR gene diversity within kākāriki karaka. These data will provide a first step towards the inclusion of TLR gene diversity in the management of kākāriki karaka.

Methods

Unless otherwise stated, I applied the same methods workflow described in Chapter 2 (Figure 2.2). The compute specifications remained the same as those outlined in Chapter 2.

Reference Genome

The kākāriki karaka reference genome was assembled by Tea Break Bioinformatics (Roger Moraga), with reads that were trimmed with TrimGalore v. 0.6.2 (Kreuger 2020) and Cutadapt v. 2.1 (Martin 2011) using an end trim quality of 30, a minimum length of 54, and the --next-seq two-colour chemistry option. Reads were assembled using both Meraculous-2D v. 2.2.5.1 (Goltsman, Ho, and Rokhsar 2017) and the MaSuRCA assembler (AV Zimin et al. 2013). BUSCO v. 3.0.1 (Simão et al. 2015) was used to assess the completeness, duplications, and missingness of the genome. While the MaSuRCA assembly (i.e., an N50 of 107.4 kb with 66,212 scaffolds > 1 kb) had fewer scaffolds and was more complete than the Meraculous assembly (i.e., an N50 of 28.5 kb with 67,046 scaffolds > 1 kb), there were several regions of poly-C/poly-G, NNNN gaps, and tandem repeats, thought to be artifacts of heterozygosity and/or issues arising from two-colour chemistry (Moraga, Personal Communication; De-Kayne et al. 2020). To help deal with these issues, the MaSurCA assembly was aligned to the Meraculous assembly using Last v. 959 (Kielbasa et al. 2011), and alignments were filtered to areas of tandem repeats or

gaps within the MaSurCA assembly. If the Meraculous assembly spanned these regions without any gaps, the sequence match from the Meraculous assembly was used instead of the MaSuRCA sequence. This process improved the completeness and quality of the final genome assembly.

TLR Identification

BLAST Alignment

The methods described in Chapter 2 to identify TLRs within the tūturuatu genome were the same ones I applied to kākāriki karaka, with changes made to the genomic resources used to identify these genes. The reference TLR sequences used for identification are listed below in Table 3.1.

Table 3.1: Reference species for toll-like receptor (TLR) identification in tūturuatu, listed alongside the taxonomic group shared with kākāriki karaka. For each species, TLR sequences used for comparison were either identified through the use of the NCBI annotation pipeline[1] or through sequencing of targeted TLRs in the laboratory.

Reference TLR Species	**Shared Taxonomic Group**	**Method of TLR Identification**
Red-crowned kākāriki (Cyanoramphus novaezelandiae)	Genera Cyanoramphus	Targeted amplification and sequencing[2]
Budgerigar (Melopsittacus undulatus)	Family Psitticulidae	NCBI annotation of genome
Kākāpō (Strigops habroptilus)	Order Psittaciformes	Targeted amplification and sequencing[2]
Zebra Finch (Taeniopygia guttata)	Class Aves	NCBI genome annotation excluding targeted amplification and sequencing of TLR4[3]

[1]The NCBI Eukaryotic Genome Annotation Pipeline, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
[2]Grueber et al. 2015, [3]Vinkler et al. 2009

I performed the reciprocal best hit test with kākāpō and zebra finch, since both have fully sequenced, well-annotated genomes. Every preliminary TLR gene identified in kākāriki karaka passed the reciprocal best hit test, mapping to the original reference sequences used to identify

them. This confirmed the match for the type of TLR gene found and supported the identification of the preliminary TLR sequences.

Then I proceeded to use BLAST to align the preliminary TLR sequences against the kākāriki karaka genome to identify false positive sequences or cross-alignment between the TLRs that might affect the alignment of resequencing data. I did not find any false positives sequences and identified only one cross-alignment between TLR1A and TLR1B, but < 50% of the gene sequence mapped between this pair. The bwa aligner default maximum mismatch value of 4% for read alignments was sufficient to prevent the cross-alignment of these TLRs.

TLR Protein Analysis

TLR protein analysis was the same as described in Chapter 2, including the use of ORF Finder, LLR Finder, and SMART as tools to investigate and refine protein structure.

Sample Extraction and Resequencing

A total of 48 individuals were resequenced as representatives of the captive population in a previous aligned study (Galla et al. 2019; Galla et al. 2020; Stephanie Galla unpublished data). Samples were extracted and sequenced in two different batches: the first batch (n=37) included eight captive families and one wild individual brought in as an egg from the Poulter, and the second (n=11) was composed of six individuals brought as eggs produced from pairs of captive females and wild males in the Poulter, four individuals brought from the translocated population on Oruawairua/Blumine, and one captive-born individual. DOC and Isaac Conservation and Wildlife Trust staff collected the blood, feather, and tissue samples during routine health checks. Samples were stored at -80°C until extraction. High quantity and quality DNA was extracted using a tailored lithium chloride extraction method (Galla et al, 2019). Extractions were assessed for quality by running 2 μl of DNA on a 2% agarose gel. A Qubit® 2.0 Fluorometer (Fisher Scientific) was used for DNA quantification.

Libraries for both batches were prepared at IKMB with Nextera™ DNA Flex Library Prep Kit used according to the manufacturer's specifications and sequenced across one lane of an Illumina Novaseq™ 6,000 at a coverage depth of approximately 10x (Galla et al, 2020).

TLR Characterisation

Population Resequencing Alignment and Analysis

I performed an alignment of the resequencing data from the captive population to the kākāriki karaka reference genome, which was assembled as described above. The trimming specifications for files were the same as in Chapter 2, because both the kākāriki karaka reference genome and resequencing data were sequenced using two-colour chemistry. I used the custom alignment script (Appendix B) for alignment of resequencing data to the reference genome and the alignment process remained the same.

I filtered the resulting population vcf file with VCFtools and parameters for filtering that had been determined in a previous whole genome alignment of this species (Galla 2019). The parameters used for filtering were: maf > 0.05, Phred-score (quality) > 20, max-missingness = 0.90, minimum depth > 5, and maximum depth < 200. I also used bcftools to filter for strand bias using the parameter of strand-bias-adjusted Phred score, SP < 60. Hardy-Weinberg equilibrium filtering was not used because the birds within the study population are in family groups and this violates the assumptions of random mating. I found that initial filtering produced very few SNPs at the TLR regions, with the majority having a SNP diversity of zero.

After this first attempt at filtering, I used vcftools to analyse the vcf for depth per site and per individual and missingness per site and per individual at TLR genes. Depth per site is a measure of how many unique reads from a population dataset map to that site, and depth per individual is a measure of how many unique reads from one individual map to all the variable sites within the region(s) of interest. Missingness per site is a percentage of how many individual genotypes are missing at each variable site out of the total population, and missingness per

individual is a percentage of how many sites an individual is missing a genotype for out of the total variable sites within the region(s) of interest.

Reviewing these statistics, I found that there were three individuals with very low depth (~1x) and intermediate to high missingness (0.12-0.55) within the TLR genes. I hypothesised that these individuals may be bringing down the average value for depth and increasing the value for missingness at TLR SNP sites, so that most SNPs were filtered out completely.

To test this, I filtered the vcf file to remove these three individuals with low depth and high missingness and created a new vcf file. Then, I filtered this file with the same parameters in VCFtools and BCFtools as described above, to see whether removing these individuals changed the resulting SNP diversity. After filtering, there were more SNPs within the TLR gene regions, so it is likely that including these low depth, high missingness individuals in the analysis was masking true diversity at these genes. I decided to use the new vcf file with the removed individuals for final TLR characterisation, which reduced the sample size from n=48 to n=45.

The removal of one individual had minimal impact on the population analysis, because this bird is from the captive population and has two half-siblings, so at least some of the genetic variation contributed by this individual is likely to be represented. Unfortunately, the other two individuals I removed were brought into the captive population, one from the translocated population on Oruawairua/Blumine and the other from a captive x wild pairing in the Poulter. Since I do not have detailed information about the parentage of these birds, it is unclear whether the SNP diversity and alleles from the individuals are still represented within the remaining population. Nevertheless, there were still three individuals from the translocated population and three individuals from captive x wild pairings in the Poulter, so the TLR gene diversity from these source populations was represented in the analysis to some extent.

As in Chapter 2, I analysed final SNPs to determine whether they made synonymous or nonsynonymous changes to TLR protein products. Haplotype construction and analysis within DNAsp followed the same method as described in Chapter 2. However, when I conducted an ad-

hoc analysis of inheritance of these haplotypes using relatedness information from the pedigree, I was not able to confirm them. In several parent-offspring groups (trios), haplotypes assigned to the offspring were unique in comparison to parental haplotypes, even when recombination was considered. These issues persisted, even when I restricted analyses to the LRR binding domain only. Therefore, it is unlikely that these data accurately reflect the true number of haplotypes. Given this, I was unable to compare haplotypes between captive birds and birds brought as eggs to supplement the population, and so could not determine whether there were unique haplotypes brought by individuals outside captivity.

Results

I identified six out of ten of the avian TLRs within kākāriki karaka (TLR1A, TLR2A, TLR3, TLR4, TLR5, and TLR7). The protein schematic structure for each of these TLRs is shown in Figure 3.1. An additional two TLRs (TLR1B and TLR2B), were partially identified, but were excluded from further analysis as I was unable to find the whole TLR sequence for either. There was no evidence of a duplication of TLR7, which is consistent with the current understanding of how this duplication segregates within the avian family tree (Velová et al. 2018).

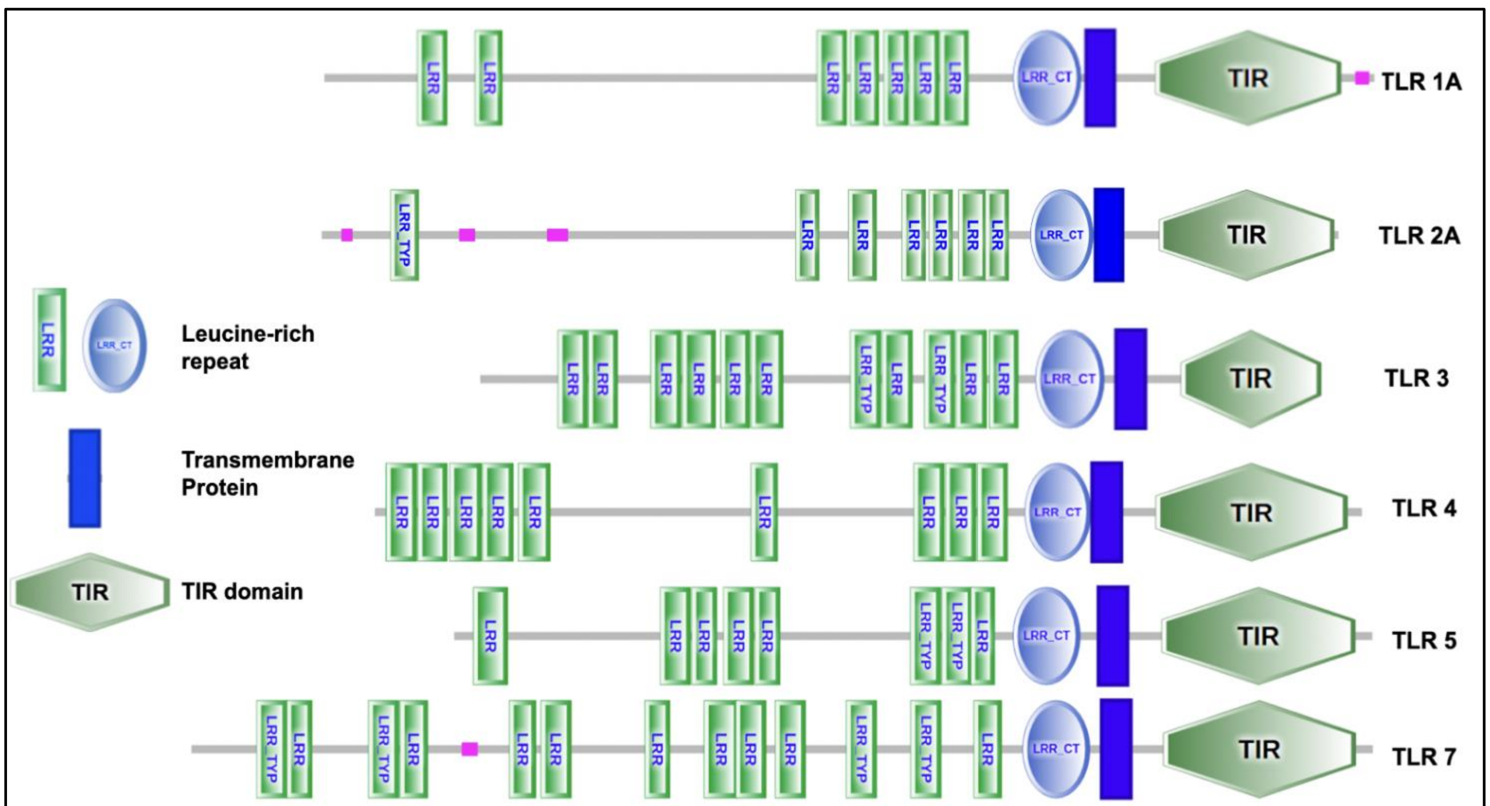

Figure 3.2: SMART protein schematics for each kākāriki karaka toll-like receptor identified (Available via license: CC BY 2.0). Key shows the visual representation for each protein domain.

There were 36 SNPs total within all kākāriki karaka TLR genes (Table 3.2). As in Chapter 2, the SNPs were unevenly distributed within the TLRs: two TLR genes (TLR1A, TLR4) with 15 and 10 SNPs, respectively, have a majority of the SNP diversity, two genes (TLR2A, TLR5) have 5 and 6 SNPs, respectively, and the remaining two genes (TLR3, TLR7) are monomorphic. Out of the total SNPs 14 were non-synonymous, and the remaining 22 were synonymous. Of the 36 SNPs, 25 were found within the LRR binding domain and 11 SNPs were found in other regions of the TLR protein. Of the former, fourteen were non-synonymous and eight were synonymous; of the latter, eight were synonymous and three were non-synonymous.

Table 3.2: Kākāriki karaka toll-like receptor sequence specifics and gene diversity statistics for captive kākāriki karaka population (n=45). Abbreviations: syn:nsyn= ratio of synonymous to non-synonymous SNPs, bp(aa)=base pairs (amino acids).

Toll-like Receptor	**Specificity[1]**	**SNPs (syn:nsyn)**	**Length bp (aa)**
TLR 1A	Bacteria, liproproteins	15 (8:7)	2484 (827)
TLR 2A	Bacteria, triacylated liproproteins	5 (1:4)	2612 (869)
TLR 3	dsRNA	0	1926 (638)
TLR 4	LPS, gram negative bacteria	10 (8:2)	2261 (752)
TLR 5	Flagellin	6 (5:1)	2113 (702)
TLR 7	ssRNA	0	2700 (899)

[1]Alcaide and Edwards 2011

In contrast to the low SNP diversity, I observed a total of 47 haplotypes across all TLR genes and haplotype diversity was also high (mean = 0.746). While this observation may be accurate it did not reflect what is observed in other species, especially in comparison to other threatened species, which have a low number of haplotypes as well as SNPs. Also, when I

analysed these haplotypes with respect to pedigree relatedness, I found that they do not follow patterns of normal inheritance. Therefore, I will not report haplotypes here or the associated statistics of nucleotide diversity, haplotype diversity, and Tajima's D.

The species of least concern chosen for comparison to kākāriki karaka were the lesser kestrel and house finch, because there were no species that have been analysed and are more closely related. The red-crowned parakeet (red-crowned kākāriki), orange-bellied parrot, and Stewart Island robin were chosen as the threatened species for comparison. The red-crowned kākāriki is in the same genera as kākāriki karaka and the orange-bellied parrot is in the same family as kākāriki karaka, but some genes were not analysed within these species, so the Stewart Island robin was also included for comparison. As in Chapter 2, I report only the SNPs within the LRR binding domain for each TLR in kākāriki karaka, because only this region was analysed in comparison species. I also did not include haplotype numbers for kākāriki karaka or any comparison species, given the issues discussed above. The resulting comparison is shown in Table 3.3.

Table 3.3: Comparison of kākāriki karaka TLR gene diversity to species and populations with different threat statuses. The species are coloured by threat status, and when possible, these statuses come from the IUCN Red List. The study where these species were analysed is cited at the bottom of the table. Abbreviations: n= number of samples, bp(aa)=base pairs (amino acids), h=number of inferred haplotypes.

TLR Locus	Species	n	Length (bp)	Conservation Status	# SNPs
TLR 1A	Kākāriki karaka (*Cyanoramphus malherbi)*	45	1815	Critically endangered	9
	Orange-bellied Parrot (*Neophema chrysogaster*)[1]	20	924	Critically endangered	4
	Red-Crowned kākāriki (*Cyanoramphus novaezelandia)*[2]	18	1165	Least Concern*	14
	Stewart Island Robin(*Petroica australis rakiura)*[3]	10	1166	Least Concern**	2
	Lesser Kestrel (*Falco naumanni*)[4]	8	1163	Least Concern	19
	House Finch (*Haemorhous mexicanus)*[4]	51	1161	Least Concern	44
TLR 2A	Kākāriki karaka (*Cyanoramphus malherbi)*	45	1942	Critically endangered	5
	Stewart Island Robin(*Petroica australis rakiura)*[3]	10	1034	Least Concern**	1
	Lesser Kestrel (*Falco naumanni*)[4]	8	543	Least Concern	6
	House Finch (*Haemorhous mexicanus)*[4]	8	560	Least Concern	13

TLR 3	Kākāriki karaka (*Cyanoramphus malherbi)*	45	1467	Critically endangered	0
	Orange-bellied Parrot (*Neophema chrysogaster*)	93	915	Critically endangered	1
	Stewart Island Robin(*Petroica australis rakiura)*[3]	10	1034	Least Concern**	0
	Lesser Kestrel (*Falco naumanni*)[4]	8	1160	Least Concern	1
	House Finch (*Haemorhous mexicanus)*[4]	8	952	Least Concern	11
TLR 4	Kākāriki karaka (*Cyanoramphus malherbi)*	45	1701	Critically endangered	5
	Orange-bellied Parrot (*Neophema chrysogaster*)[1]	20	738	Critically endangered	3
	Red-Crowned kākāriki (*Cyanoramphus novaezelandia)*[2]	20	849	Least Concern*	8
	Stewart Island Robin(*Petroica australis rakiura)*[3]	10	1034	Least Concern**	4
	Lesser Kestrel (*Falco naumanni*)[4]	8	818	Least Concern	6
	House Finch (*Haemorhous mexicanus)*[4]	8	789	Least Concern	16
TLR 5	Kākāriki karaka (*Cyanoramphus malherbi)*	45	1514	Critically endangered	6
	Orange-bellied Parrot (*Neophema chrysogaster*)	49	864	Critically endangered	0
	Red-Crowned kākāriki (*Cyanoramphus novaezelandia)*[2]	20	961	Least Concern*	7
	Stewart Island Robin	10	1229	Least Concern**	2
	Lesser Kestrel (*Falco naumanni*)[4]	8	1265	Least Concern	20
	House Finch (*Haemorhous mexicanus)*[4]	8	951	Least Concern	2
TLR 7	Kākāriki karaka (*Cyanoramphus malherbi)*	45	2049	Critically endangered	0
	Orange-bellied Parrot (*Neophema chrysogaster*)	20	579	Critically endangered	0
	Red-Crowned kākāriki (*Cyanoramphus novaezelandia)*[2]	18	730	Least Concern*	1
	Stewart Island Robin (*Petroica australis rakiura)*[3]	10	618	Least Concern**	3
	Lesser Kestrel (*Falco naumanni*)[4]	8	869	Least Concern	3
	House Finch (*Haemorhous mexicanus)*[4]	8	982	Least Concern	27

Total SNPs (syn:nsyn)	Kākāriki karaka	Orange-bellied Parrot	Red-crowned kakariki	Stewart Island Robin	Lesser Kestrel	House Finch
	25 (14:8)	7 (1:6)***	30 (16:15[a])***	12 (2:10)	55 (32:23)	113 (63:50)

References: Morrison et al 2020, [2]Grueber et al 2015, [3]Grueber et al 2012, [4]Alcaide and Edwards 2011. *Listed as species of least concern, but population analysed is small and threatened Grueber et al 2015. **Listed as a species of least concern, but the population used in this study is small and isolated Grueber et al. 2012. ***Data are incomplete for these species, and totals reflect 5 TLRs in orange-bellied parrot and 4 TLRs in red-crowned kākāriki. [a]Triallelic SNPs were observed at TLR4 in red-crowned kākāriki

For the reported TLR genes in red-crowned kākāriki, the SNP diversity was similar to that within TLR genes of kākāriki karaka. For example in genes where SNP diversity was high in red-crowned kākāriki, it was also high in kākāriki karaka. The total number of SNPs in the three threatened species were similar and were lower than the total SNPs in the species of least

concern. Kākāriki karaka have more synonymous than non-synonymous SNPs, which is similar to the ratio of synonymous to non-synonymous SNPs within the species of least concern.

Discussion

From a methodological standpoint, this chapter revealed some technical challenges. In particular, I found that individuals with low-depth and high-missingness may mask variation within a population. While removing these individuals with low-depth and high-missingness resulted in some data loss, it did allow me to characterise TLR gene diversity in the remaining individuals. However, I was not able to accurately haplotype the remaining individuals. It is unclear why the phased haplotypes I found within this population did not conform to expected inheritance models. Since SNP diversity is characterised on a population level, the variable sites are likely to be accurate, because they hold up across the whole population. In contrast, haplotyping requires information about the nucleotide identity at a variable site in any one individual, and there may not be enough reads within an individual alignment to accurately determine the identity of this nucleotide. While I did set the genotype of low-depth and high strand-bias sites to missing and removed all haplotypes that had missing sites, there may have been additional issues within individual SNP sites that I did not consider. Using a different genotyper may result in more accurate haplotypes and this will be considered in future research.

The issues that I experienced when applying this bioinformatic approach to kākāriki karaka have important implications for its future use. One is that when characterising functional diversity at a few specific genes, it is important that the depth is high and missingness is low within these regions. In a conservation context, if having data for a few individuals is especially important, it may be worth sequencing these individuals at a higher depth. For example, in the case of kākāriki karaka, I had to remove some of the wild- or translocated-sourced individuals, and I may not have captured all of the TLR gene diversity that has been brought into the captive population. So, in hindsight, given the importance of these individuals, they could have been sequenced at 20x

rather than 10x depth. Another caveat is that this method is adequate for characterisation of SNP diversity at the population level, but may not be sufficient for identifying haplotypes at an individual level. This method could be a first step to determine variation within the TLR genes at a population level, but if knowing individual haplotypes is essential to answer a particular research question, then targeted amplification and sequencing of TLR genes provides another option for haplotyping individuals.

Despite these challenges, I was able to characterise TLR SNP diversity in kākāriki karaka and found that the overall SNP diversity is low and similar to that of other threatened species with small populations. The kākāriki karaka captive population has been periodically augmented with eggs from multiple source populations in an attempt to minimise the loss of neutral genetic diversity, but it is unclear what has happened in respect to TLR gene diversity. While overall TLR SNP diversity within captivity is low, it is possible that it would have been even lower than what I measured if there had not been previous augmentation. For example, if TLR gene diversity was relatively high in the wild despite past reductions in population size, or if TLR gene diversity was relatively low, but the wild population harboured different TLR alleles than the captive population, then periodic augmentation may have increased TLR SNP diversity in the captive population. Indeed, preliminary evidence indicates that individuals recently sourced from the Poutler and Oruawairua/Blumine will increase genome-wide neutral genetic diversity of the captive population if they become established breeders (DOC RG, unpublished data). On the other hand, augmentation may not have increased TLR gene diversity in the captive population at all. For example, it is unclear how much TLR gene diversity was present when birds were translocated to Oruawairua/Blumine, and because the translocated population has not been augmented since 2012 it is possible that genetic drift has contributed to further loss of TLR gene diversity and that the alleles remaining are already well-presented within the captive population.

Regardless, I found that SNPs in TLR genes within the captive population are not distributed equally. There are some genes with a high degree of SNP diversity while others are

monomorphic. The loci with high SNP diversity were TLR1A, TLR2A, TLR4, and TLR5, which are mostly specific for bacteria. Many TLRs have been implicated in recognition of and activation of the immune system against fungal pathogens, including TLR1A, TLR2A, and TLR4 (Cheng et al. 2020; Leishangthem et al. 2015). The higher genetic diversity within TLR1A, TLR2A, and TLR4 genes may result in a stronger immune response to fungal infections like Aspergillosis. Alternatively, this finding may reflect the results in previous studies in humans and birds that there is higher SNP diversity within non-viral sensing TLRs (Barreiro et al. 2009; Grueber 2015). There tends to be higher polymorphism within these non-viral TLRs, because they are redundant, meaning multiple kinds of TLRs recognise the same bacterial and fungal pathogens (Barreiro et al. 2009). I observed this pattern within kākāriki karaka because the genetic diversity of non-viral sensing TLRs was high and the two viral-sensing TLRs were monomorphic (TLR3 and TLR7).

Regardless of the cause, the lack of SNP diversity within TLR3 and TLR7 genes means that the captive population may be more susceptible to viruses like BFDV. The study of red-crowned kākāriki on Little Barrier Island found that one variant of TLR3 may have been selected through exposure to BFDV (Knafler et al. 2016). In the case of kākāriki karaka, there is no possibility for such selection, and it is unknown whether the single variant within the captive population is sufficient to produce a robust immune response. It appears unlikely that this variant is the result of prior virus-induced selection since there are no records of such an outbreak in either wild or captive populations (DOC RG).

Conclusions

Overall SNP diversity within the captive kākāriki karaka population is low. While there are no current disease outbreaks within this population, the threat of pathogens only continues to increase within Aotearoa New Zealand (Alley and Gartrell 2019). The captive population of kākāriki karaka exists both to introduce young birds back into the wild and as an insurance

population (DOC RG), and so increasing immune gene diversity in advance of a disease outbreak may improve outcomes. Further analyses of TLR gene diversity within the wild and translocated populations and within all offspring of the wild x captive pairs will allow for a more complete comparison with the captive population and would add value to existing conservation management actions.

Chapter 4: Challenges associated with an incomplete reference genome and resequencing data prevent characterisation of immune gene diversity in a threatened endemic bird, kakī/black stilt

Abstract

In the third stage of this Proof of Concept, I apply the same bioinformatic approach to determine whether I can characterise TLR gene diversity in kakī/black stilt. Kakī/black stilt is a threatened bird species endemic to Aotearoa New Zealand and has a semi-wild population, which is composed of individuals within the captive breeding programme and wild eggs that are reared in captivity and released back into the wild as juveniles. I encounter a number of technical issues that prevent identification of some of the TLR genes within the reference genome and characterisation of the TLR genes I am able to identify. After examining the reference genome, I identify poly-N repeats within and nearby TLR genes, which likely prevent full identification of these TLRs. I also investigate the alignment of population resequencing data through a series of tests and conclude that widespread missingness within the resequenced population may prevent characterisation of TLR genes. These findings, combined with the results in Chapter 2 and 3, suggest this bioinformatic approach requires a complete reference genome and resequencing data to be effective.

Introduction

Kakī/Black stilt (Himantopus novaezelandiae) is a nationally critical species in Aotearoa New Zealand with a wild population of approximately 169 adults known-to-be-alive (Forsdick 2020, Department of Conservation Recovery Group). The original population of kakī spanned

both main islands but experienced a rapid decrease in population size after the introduction of mammalian predators in the 20th Century (Reed 1998). The species has also experienced declines due to the loss and degradation of their native braided river habitat (Pierce 1984). The current wild population of kakī is almost exclusively restricted to Te Manahuna/Mackenzie Basin and is intensively managed (Galla et al. 2019). A conservation breeding and rearing programme for the species was started in the 1980's, when it was estimated there were only 23 birds left in the wild (Steeves et al. 2010). The programme facilitates breeding of captive pairs and brings in eggs from the wild for captive rearing. Captive rearing improves outcomes for juveniles, because predation on eggs and chicks in the wild is much higher, and individuals are more likely to survive to the juvenile stage within captivity (DOC RG). The collective aim of these two captive programmes is to release as many juveniles as possible into the wild population (Galla et al. 2020; Overbeek et al. 2020). Therefore, in this thesis, I will refer to the entire kakī population as "semi-wild," because the vast majority of wild adults known-to-be-alive were raised in captivity (DOC RG).

As of 2020, eggs from almost all wild nests are brought into captivity (i.e. of the 169 adults known-to-be alive, 162 were captive reared, five were wild reared, and two were unknown; DOC RG). However, due to intensive predator control in Te Manahuna, there may soon be more eggs from wild nests than can be accommodated in the captive rearing programme. In future, the programme may make decisions about which individuals to prioritise for captive rearing, like preferentially bringing in eggs from pairs that include founder lineages that are not well-represented within the kakī pedigree (Figure 4.1, reproduced with permission from Galla 2019). However, it is unknown how such decisions, based on neutral genetic diversity, may influence the TLR gene diversity of the semi-wild population.

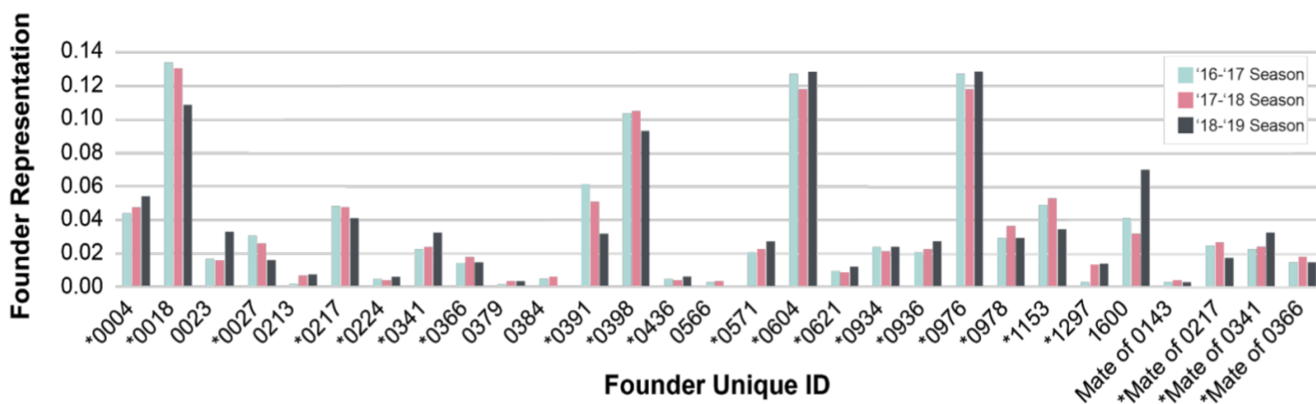

Figure 4.1: Founder representation across all breeding individuals over three breeding seasons. Founders with an asterix (*) by their ID have been captured in the captive population over the past three seasons. Reproduced with permission from Galla 2019.

I attempted to characterise TLR gene diversity within the semi-wild kakī population and to compare individuals with different founder lineages to determine whether maximising representation of founder lineages that are not well-represented in the semi-wild population may improve TLR gene diversity. However, I was only able to identify four complete TLR genes and was not able to characterise diversity at these genes. To identify which characteristics of the data may be causing these problems, I analysed the reference genome and resequencing data. I was unable to unequivocally determine the cause(s) of these underlying these issues, but the incompleteness within the reference genome and within the short-read resequencing data were contributing factors. In future, a new kakī reference genome will be assembled with PacBio long-read sequences and this will lead to a more complete assembly with longer scaffolds, which may allow for greater identification of TLRs and characterisation of TLR gene diversity.

Methods

Reference genome

The reference genome used for this method was assembled by Natalie Forsdick with reads trimmed for adapters using Trimmomatic v. 0.35 (Bolger, Lohse, M., and Usadel 2014) and for quality (Phred score > 30) using ConDeTri v2.3 (Smeds and Künstner 2011). Unlike in the previous two species, the two-colour chemistry option was not used because the reads were sequenced using four-colour chemistry. FastQC was used to assess paired-end read quality before and after cleaning, with 99.1% of trimmed reads achieving a quality Phred-score > 30 (Forsdick 2020). Multiple assemblers were used for primary assemblies and analysed for quality and completeness using BUSCO (Simão et al. 2015). The highest quality assembly with the longest scaffolds was produced by SOAPdenovo2 (Luo et al. 2012), so this was chosen as the final assembler. After initial assembly, gap filling and super-scaffolding the draft genome with the chicken genome improved contiguity (Forsdick 2020). There is an upcoming update to the kakī genome assembly being assembled with PacBio long reads and a modified trio-binning approach, but this assembly was not yet available to use for TLR characterisation (Natalie Forsdick and Roger Moraga, Personal Communication).

TLR Identification

BLAST Alignment

The method used for TLR identification with BLAST was the same as that in Chapter 2. The reference TLR sequences used for identification are listed below in Table 4.1.

Table 4.1: Reference species for toll-like receptor (TLR) identification in tūturuatu, listed alongside the taxonomic group shared with kakī. For each species, TLR sequences used for comparison were either identified through the use of the NCBI annotation pipeline[1] or through sequencing in the lab. Links to the study in which TLRs were sequenced are at the bottom of the table

Reference TLR Species	Shared Taxonomic Group	Method of TLR Identification
Killdeer (Charadrius vociferus)	Order Charadriiformes	NCBI genome annotation
Black-headed Gull (Chroicocephalus ridibundus)	Order Charadriiformes	Targeted amplification and sequencing[2]
Chicken (Gallus gallus)	Class Aves	Targeted amplification and sequencing[3]
Zebra Finch (Taeniopygia guttata)	Class Aves	NCBI genome annotation excluding targeted amplification and sequencing of TLR4[4]

[1]The NCBI Eukaryotic Genome Annotation Pipeline, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
[2]Podlaszczuk et al. 2020, [3]Yilmaz et al. 2005, [4]Vinkler et al. 2009

Since several TLRs were only partially identified within the kakī genome, I conducted further analysis of these regions by extracting the partial TLR sequence and the 1kb window around this sequence to determine whether there were any aspects of the sequence or surrounding sequences that were preventing full identification of the gene.

For the TLR sequences that I did fully identify, I conducted a false positive analysis using the same method as in Chapter 2 and performed the reciprocal best hit test using the Chicken and Zebra Finch genomes. All TLRs passed the reciprocal best hit test, and I did not identify any false positives or cross-alignment between TLRs. Then, I conducted the same protein analysis as described in Chapter 2.

Sample Extraction and Resequencing

A total of 36 individuals were chosen for resequencing to represent the most diverse lineages identified through pedigree analysis (Galla 2019). Brett Gartrell and Isaac Conservation and Wildlife Trust staff collected blood samples during routine health checks. Samples were stored at -80°C until extraction. Another member of the research team (Stephanie Galla), extracted high quantity and quality DNA using a tailored lithium chloride extraction method (Galla et al, 2019). Extractions were assessed for quality by running 2 μl of DNA on a 2% agarose gel. A Qubit® 2.0 Fluorometer (Fisher Scientific) was used for DNA quantification.

Resequencing libraries were prepared at IKMB using a TruSeqR Nano DNA Library Prep Kit, following the manufacturer's protocol, and were sequenced across 2 lanes of an Illumina HiSeq 4,000. Twenty-four of the resequenced genomes were sequenced at a high coverage depth (~50x) for an aligned study, and all others were sequenced at a lower coverage depth (~10x).

TLR Characterisation

Population Resequencing Alignment

Short-read resequencing data were aligned to the reference genome using the custom alignment script (Appendix B). The final vcf file was filtered with the following parameters: maf > 0.05, Phred-score (quality) > 20, max-missingness = 0.90, minimum depth > 5, and maximum depth < 200. However, filtering resulted in no SNPs within TLR genes. Given the results from other species, I questioned whether or not this result was an accurate representation of the TLR gene diversity in kakī or whether the absence of SNPs may be a result of issues with the genome alignment. I ran a number of quality tests to investigate what was happening with the alignment (see Table 4.2 for a summary).

Similar to the analysis of partial TLR sequences within the reference genome, I first analysed the 1 kb window around the TLR genes to determine whether there were areas of incompleteness within these neighboring sequences.

Another hypothesis for the lack of SNPs within TLR genes was that if there were reads with high heterozygosity within the resequencing data, they may not map to the reference genome. To get a greater understanding of what may be happening in the alignment process, I conducted an analysis of DNA reads that did not map during the alignment, using a script adapted from Laine et al 2019 (Appendix C). I used SAMtools to extract unmapped reads from each individual bam file, created a new bam file with these unmapped reads, and sorted the resulting bam files. Then, I converted the bam files into paired fastq files using BEDtools (Quinlan and Hall

2010). I used the SPAdes assembler (Bankevich et al. 2012) for local assemblies of unmapped reads in the fastq files. Finally, I converted the assembled fastq files into fasta files, which I concatenated so that the unmapped reads from all individuals in the dataset were contained in one fasta file.

I used the BLAST+ program to create a database from the unmapped reads fasta file, and I used BLASTn with the TLR sequences I had identified within the kakī genome as query sequences. This helped me to determine whether there were TLR sequences present within the unmapped reads, meaning that they had not aligned to the reference genome. In addition to using kakī TLR sequences as query sequences, I used all of the original reference TLR sequences that I had for identification (Table 4.1), to investigate whether there were TLR sequences within the unmapped reads that I had not been able to fully identify in the reference genome. I found two fragments that aligned to multiple TLR7 reference sequences, but did not find the full TLR7 sequence or any other TLR sequences within the unmapped reads.

My next test was repeating the population resequencing alignment with NextGENMap (Sedlazeck, Rescheneder, and von Haeseler 2013), which is more permissive aligner, and I included settings for high sensitivity, to prioritise mapping reads over matching the identity of bases in reads to the identity of bases in region of alignment. I conducted this alignment to test whether additional, more heterozygous reads might map to TLR genes. As a test of the new aligner, I aligned a few individual fastq files. After alignment of fastq files to the reference genome, I input the resulting bam files into Geneious, which allowed me to visualise reads within each bam file and to determine whether there were additional reads mapping to TLR genes, without having to run a whole alignment pipeline. I did not find additional reads mapping within TLR genes, even when I increased sensitivity settings to --very-high.

Then, I quantified a read depth across the whole genome compared it to read depth at TLR genes. This analysis helped me to find whether read mapping rates were the same within TLR genes as they were genome-wide. For each individual bam file, I used SamTools to count

reads at every 1kb window along a contig and these counts were output into a csv. These data were input into R, and I used ggplot to visualise the depth for each individual within every chromosome. I compared read depth counts within TLR genes to the read depths genome-wide, and found there was no significant difference.

I then analysed the unfiltered vcf file and found SNPs within some TLR genes (TLR2A had 11 SNPs, TLR5 had 6 SNPs, TLR3 and TLR4 were monomorphic). I then conducted a filtering trial for the population vcf file to find which filters were causing the greatest decrease in SNP diversity within TLR genes. I used VCFtools to filter the file using each parameter individually (eg. minQ > 20) and then I evaluated the resulting SNPS in each filtered file (see Appendix D). I found that the filter that caused the greatest decrease in SNP diversity was missingness, because all TLR genes were monomorphic after I applied this filter.

I then conducted a second filtering trial with missingness, incrementally increasing the percent of missing data by 10% and examining the resulting SNPs (see Appendix D). I found that increasing missingness increased the resulting SNPs within TLR genes (see Results). I also quantified individual and site missingness and found widespread intermediate to high missingness within the population. While these results reveal important characteristics of the population resequencing data and alignment, I do not feel confident saying that they represent the SNP diversity at TLR genes without further data.

Table 4.2: Summary of bioinformatic tools used for quality tests of the population resequencing alignment for kakī. See text for details. Abbreviations: TLR= toll-like receptor, SNP=single nucleotide polymorphism. All remaining abbreviations refer to names of bioinformatic tools utilised.

Bioinformatic Tool(s)	**Purpose**	**Results**
SAMtools	To extract 1 kb windows around TLR sequence and search for incompleteness (Ns)	There were no N’s discovered in the areas near TLR genes
SAMtools, BEDtools, SPAdes	To identify if there were TLR sequences in unmapped reads	A few fragments indicate there may be pieces of TLR7 in unmapped reads, but no other TLRs were detected

NextGenMap, Geneious	To realign resequencing data with more permissive settings that encourage mapping of more heterozygous reads	Examination of bam files did not reveal additional SNPs due to more reads mapping
SAMtools, Bedtools	To count reads genome-wide and within TLR regions to compare mapped reads to compare overall depth to depth at TLR genes	Read depth within TLR regions is comparable to that along the whole genome.
VCFtools	To conduct filtering trials for missingness, given the widespread degree missingness within the population	At higher missingness values, there are SNPs within TLR2A and TLR5, but I do not have confidence in the accuracy of these results.

Results

I identified three partial TLR sequences (TLR1A, TLR1B, TLR2B). When I examined the sequences and the 1kb neighbor sequences on either side of them, I found that there were poly-Ns within the partial TLR sequences or within close neighboring sequences. It is likely that the incomplete TLRs are not well-assembled in the kakī reference genome and that incompleteness prevents full identification of some TLR genes. For example, within the partial sequence of TLR1A there was a large poly-N repeat at the start of the sequence (Figure 4.2).

TAAAGTGATGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCGGCA

Figure 4.2: The first 118 bp of the partial TLR1A sequence shows a 100 bp poly-N repeat

I was able to identify four complete TLRs in the kakī reference genome (TLR 2A, TLR3, TLR4, TLR5) (Figure 4.3), and the respective length of each TLR is as follows: 2550 bp (848 aa), 1621 bp (535 aa), 2259 bp (751 aa), and 2573 bp (847 aa).

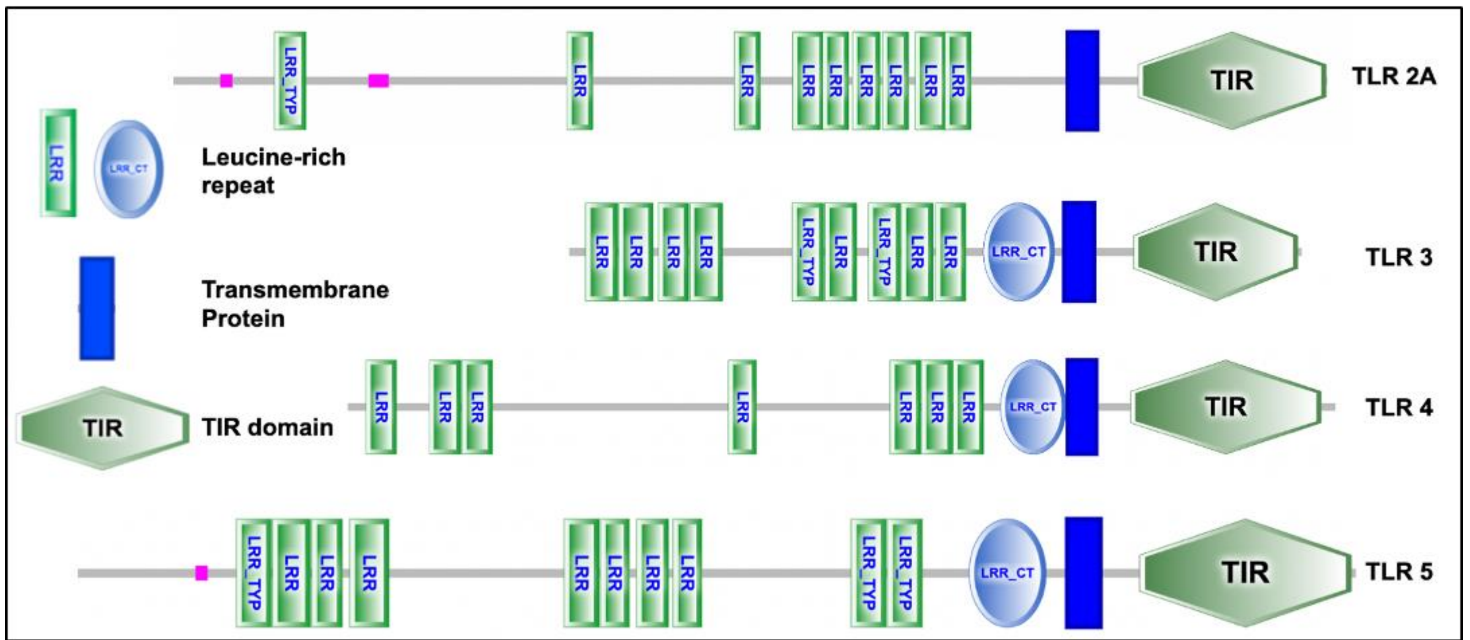

Figure 4.3. SMART protein schematics for each kakī toll-like receptor identified (Available via license: CC BY 2.0). Key shows the visual representation for each protein domain.

Within the four complete TLR genes, there were no SNPs revealed from the population resequencing alignment. It was unclear whether this result was due to true homozygosity within the kakī TLR genes or whether there may be additional data quality issues within either the reference genome or the resequencing data that may be preventing identification of SNPs within these TLR genes.

To investigate incompleteness in the reference genome, I analysed each TLR sequence and the 1kb regions surrounding them. However, I did not find any poly-N repeats or other artifacts of incompleteness in these regions, as was seen within and nearby the TLR genes that were not fully identified.

To analyse the alignment of the resequencing data to the reference genome, I conducted an unmapped reads analysis to investigate whether there were TLR sequences in the resequencing reads that were not aligning to the reference genome. No complete TLR sequences were found within these reads, but I was able to find some fragments that mapped to TLR7 sequences from multiple reference species.

Then I tested whether I could get more reads to align with a more permissive aligner and settings, but no SNPs resulted. To further investigate whether there were fewer reads mapping at TLR genes, I quantified the amount of reads at TLR regions and compared this to the amount of reads mapping genome-wide, but did not find a significant difference.

I returned to the original vcf file and conducted a preliminary filtering analysis. I found that filtering for missingness led to the greatest decrease in SNPs. I proceeded to do a series of missingness filtering trials to evaluate how changing the percent of missing data allowed at variable sites would influence the resulting number of SNPs. When I decreased missingness from 0.9 (10% missing data) to 0.8 (20% missing data), it resulted in 5 SNPs within TLR2A and 5 SNPs within TLR5. As I further decreased the value for missingness to 0.7 (30% missing data), it resulted in 9 SNPs within TLR2A and TLR5 still had 5 SNPS. Decreasing missingness beyond 0.7 made no difference until I filtered for a missingness of 0.2 (80% missing data), and it resulted in 11 SNPs within TLR2A and 6 SNPs within TLR5, which were the original SNP values in the unfiltered vcf file.

Once I found SNPs at the lower missingness filter, I was also able to quantify the individual and site missingness for the population at these variable sites in the TLR genes. I found a widespread degree of missingness within the population. Almost half of the individuals (n=16) had an intermediate to high degree of missingness (0.2-0.8), meaning that they were missing genotypes at 20-80% of variable sites. Unlike in Chapter 3, I could not remove the high missingness individuals without losing a significant amount of data.

At this stage, I choose to refrain from any additional analyses of these resequencing data. I was not confident in the accuracy of the SNP diversity produced with a lower missingness value, because a large proportion of the population lacked genotypes at one or several of the variable sites.

Discussion

When I applied this bioinformatic approach to kakī, I had limited success in identifying TLR genes in the reference genome, and I was not able to characterise TLR gene diversity within the semi-wild population. I found partial sequences for three TLR genes and identified four complete TLR genes, but was not able to characterise diversity in the genes I fully identified. I applied multiple bioinformatic methods to investigate what aspects of the genomic resources may be causing these issues. There are likely multiple reasons–which are not mutually exclusive–for these issues, but I was able to identify that both incompleteness of the reference genome and high missingness within the resequencing data presented significant barriers to TLR identification and characterisation in kakī.

The poly-N repeats within and nearby some partial TLR sequences likely prevented identification of the complete TLR sequences. These results reflect the BUSCO analysis of the genome, which shows this assembly has a relatively high %N of 3% (Forsdick 2020). In contrast, the TLR genes I was able to fully identify did not have N's either within the sequences or in neighboring sequences which supports the theory that incompleteness is what caused problems for the partially identified TLRs. Recent PacBio long reads and a modified trio-binning approach will likely result in increased quality of the kakī reference genome, such that gaps within TLR genes will be filled (Natalie Forsdick and Roger Moraga, Personal Communication).

While it was relatively straightforward to determine why I could not identify TLRs, it was more difficult to determine what may be interfering with characterisation of TLR genes. Since the complete TLR genes did not show the same pattern of N's within or nearby the sequences, poly-N gaps were likely not the cause. I subsequently investigated the issue through a number of tests, which I describe below.

While I did not find any full TLRs within the unmapped reads from bam files, I did find two fragments that aligned to reference TLR7 sequences. Since I was not able to identify TLR7 within the kakī reference genome, this result is further evidence that TLR7 may be incomplete, split between scaffolds, or otherwise not fully resolved within the reference genome. There may be

additional TLR sequences within the unmapped reads that were not assembled or are too small or incomplete to map to the TLR gene databases, and so I was unable to identify them. It is also possible that there were no other TLR genes within the unmapped reads.

Realignment of the reads with a more permissiver aligner with high sensitivity settings did not result in any SNPs within the TLR genes. I then compared the number of mapped reads across the whole genome to reads within TLR genes and found they were not significantly different. This suggests that my inability to characterise TLR gene diversity was not caused by a lack of reads mapping to TLR genes.

When I filtered SNPs with a lower percent missingness, I found some SNPs within two TLR genes (TLR2A and TLR5). However, many individuals within the population had an intermediate to high degree of missingness at these variable sites. This indicates that there may be reads missing at the individual level that were not detected at the population level. The resequencing reads might have similar sites with poly-N gaps within or nearby TLR genes, as was observed in the reference genome, which would prevent alignment and may not have been easily detected within the unmapped reads analysis. It is also possible that these individuals have a high degree of missingness across the whole genome, and so the comparison of genome-wide read counts to TLR gene read counts would remain consistent, because the same amount of reads are missing genome-wide.

Missingness within individuals may be caused by a relatively higher degree of heterozygosity within these individuals, since they were chosen to represent the most diverse lineages within the pedigree. If individuals in the population are highly heterozygous, resequencing reads may not map as well if the reference genome lacks the genomic variation that exists in resequencing data. If heterozygosity contributed to alignment issues, it implies that this approach may not be as successful for species with highly heterozygous populations and may work better in threatened species because of the low heterozygosity within their populations.

Alternatively, in the case of some individuals, high missingness may have been caused by the lower coverage sequencing carried out for a separate study. Of the 16 individuals with some degree of missingness, ten were sequenced at a lower coverage. Still, low depth of coverage does not completely explain the high missingness individuals, because the remaining six individuals with high missingness were sequenced at a high depth of coverage.

Regardless of the cause, the high degree of missingness within the population prevented characterisation of diversity within TLR genes. Once the new long-read genome becomes available, this approach can be tested again to see whether it improves outcomes for characterising TLR gene diversity, as well as for identifying TLRs within the reference genome. Alternatively, if bioinformatic identification and characterisation continues to present a challenge, then targeted lab sequencing of TLR loci can be utilised for characterisation of TLR gene diversity in the semi-wild kakī population.

Conclusions

While the attempt to identify and characterise TLR genes within kakī was largely unsuccessful, it has important implications for employing this bioinformatic method in other species. A complete reference genome is essential for identification of TLR genes and low missingness of resequencing data is important for characterisation of these genes. A shift towards long-read sequencing may improve completeness of the reference genome and may reduce missingness and improve quality of the population resequencing data.

Chapter 5: General Discussion and Future Directions

In this thesis, I leveraged existing genomic resources and developed a novel bioinformatic approach to identify TLR genes and characterise TLR gene diversity in tūturuatu, kākāriki karaka, and kakī as a Proof of Concept. These data provide a critical first step towards using TLR gene diversity to inform conservation actions for three threatened Aotearoa New Zealand endemics and towards the use of this approach to evaluate TLR gene diversity and inform conservation actions for other threatened bird species.

For tūturautu (Chapter 2), I was able to identify nine TLRs and characterise all TLRs that I identified. I found that the captive population has a low TLR gene diversity overall, similar to other threatened bird species. The SNP diversity within the population was not equally distributed within TLR genes and, in particular, low diversity within TLR21 may be associated with the impaired immune response to avian pox infection and vaccination observed in captive individuals. Future characterisation of TLR genes within wild tūturautu will inform whether bringing individuals from the wild population into captivity may improve TLR gene diversity.

For kākāriki karaka (Chapter 3), I was able to identify seven TLR genes and characterise SNP diversity of all TLRs I identified, but only after I removed individuals with low depth and high missingness. Despite periodic augmentation, the overall TLR SNP diversity within the captive population was low and similar to that of other threatened bird species. I was not able to phase haplotypes and so could not determine whether supplementation introduces novel TLR alleles into the captive population, but the low SNP diversity I found suggests that the degree to which TLR gene diversity may increase from supplementation depends on the existing genetic diversity within source populations. Future characterisation of the wild and translocated populations, and of wild x captive offspring, would allow for a full comparison of these sources with the captive population and would inform if bringing in eggs from outside captivity increases TLR gene diversity.

For kakī (Chapter 3), I was only able to identify four TLR genes and was not able to characterise TLR gene diversity at all. I identified incompleteness within the reference genome and high missigness population resequencing data that may have contributed to these problems. An upcoming assembly of a long-read reference genome for kakī will likely improve TLR gene identification and may improve characterisation of TLR gene diversity for the semi-wild kakī population.

Overall, this Proof of Concept confirmed that TLR genes can be identified and TLR gene diversity can be characterised using a bioinformatic approach, but the degree to which diversity can be characterised is dependent on the quality of genomic resources for a species. As discussed in Chapter 4, if the reference genome has a relatively high degree of incompleteness, gaps within or around TLR sequences may prevent identification of TLR genes. Also, the quality of resequencing data is important, because if there are individuals with low depth or high missingness in a dataset (Chapter 3, Chapter 4), they may mask true TLR gene diversity within the population. While including individuals with low quality reads may not be problematic when characterising genome-wide neutral genetic diversity using thousands of SNPs, to characterise diversity at specific genes it is necessary to include only those individuals with high quality reads, especially if the goal is to generate individual haplotypes.

While these are important technical considerations for the broad applicability of this approach to other bird species, they do not undermine its value. Indeed, if genomic resources are already developed for a species of interest, these can be readily leveraged to provide a first look at TLR genes and TLR gene diversity. Further, even if TLR identification and characterisation is incomplete or wholly unsuccessful, a bioinformatic approach could produce useful information about TLR regions that can inform targeted sequencing of TLR genes (e.g., even partial information about TLR sequences could inform primer design).

Compared to identifying and characterising TLRs using traditional lab-based approaches, the novel bioinformatic approach developed in this thesis is much cheaper, and may be faster

and more efficient than targeted amplification and sequencing. For example, this thesis required limited fiscal spend because genomic resources were already generated, access to the compute cluster was a free university resource (though this may not be the case for all who are interested in using this bioinformatic approach), and the bioinformatic tools I used are freely available. Also, I was able to identify and characterise the same number of or more TLR genes in tūturautu and kākāriki karaka as compared to similar studies done using traditional lab-based approaches (Grueber et al. 2015; Dalton et al. 2016; Morrison et al. 2020). If characterisation with this bioinformatic approach is successful, information about TLR gene locations within the reference genome allows for future characterisation of additional individuals or other populations of the same species. In sum, while targeted sequencing remains a useful method for assessing functional diversity, the bioinformatic approach described here provides another option for evaluating TLR gene diversity in species that have existing whole genome sequences and population resequencing data.

Future Directions

In the following sections I outline future directions for each species and highlight how new data and genomic technologies can be integrated into existing analyses. I also provide recommendations for applying this approach to other bird species. Then I describe new genomic technologies that may be integrated into this approach in the future

Tūturuatu

I identified almost all TLRs genes within the tūturuatu reference genome, with the exception of the complete TLR15 gene and a possible duplication of TLR7. Long-read sequencing of the genome would likely improve identification of these TLRs and allow greater characterisation of TLR gene diversity within tūturuatu. In any case, the TLR genes I did identify within the tūturuatu

reference genome and the bioinformatic method described in this thesis will enable all future characterisation of TLR gene diversity in the wild population.

Analysis of the TLR gene diversity of wild individuals brought to supplement the captive population in December 2020 and of TLR gene diversity of the resulting offspring from wild x wild pairs will directly evaluate whether supplementation leads to increased TLR gene diversity. These data will be paired with results from vaccination trials with wild chicks and subsequent offspring to evaluate whether there is an increase in immune response to vaccination. Together, these data will reveal whether the wild tūturuatu population has a higher TLR gene diversity and an associated increase in immune function in comparison to the captive population.

While TLRs play an important role in recognition of pathogens and activation of the immune system, there are other innate and adaptive immune cells and pathways that are necessary for a fast and effective response against pathogens like poxvirus. Further analyses of diversity within other immune genes like interleukin receptors or MHC would give greater insight into the immunocompetence of captive tūturuatu and how management actions may improve immune gene diversity on a wider scale. Also, post-vaccine analysis of other products like cytokines and tumor necrosis factor (TNF), from the innate immune response, or memory T and B cells, from the adaptive immune response, would allow for a greater understanding of which immune cells and receptors are activated and responding to vaccination. Ideally, a gene expression study within an immunocompetent, related bird infected with poxvirus would provide data about which immune genes are activated during disease progression and which are most important to overcome poxvirus infection. These investigations would allow for a targeted genetic analysis of immune genes that are most important for a robust response to poxvirus and would inform conservation actions to improve the diversity of these immune genes within captive tūturuatu.

Kākāriki karaka

There were several TLRs that I was not able to fully identify within the kākāriki karaka reference genome (TLR1B, TLR2B, TLR15, and TLR21). A long-read reference genome may improve identification of these genes, in particular of TLR1B and TLR2B, which I partially identified in the reference genome. The TLR genes that I fully identified within the reference genome will enable further characterisation of TLR gene diversity in kākāriki karaka.

In future, genomic data from the wild and translocated populations and wild x captive offspring should be analysed to determine whether existing genetic diversity in these source populations has been captured in the captive population through past and ongoing augmentation. If such future samples are collected, it may be useful to do a preliminary screening of resequencing data for depth and missingness, to inform whether repeating library resequencing may improve data quality and prevent the issues that I experienced in this thesis. Ideally, additional samples would be sequenced with long-read sequencing technology to improve the completeness of resequencing data and quality of individual alignements. However, if individual haplotypes still cannot be identified with a bioinformatic approach, then targeted sequencing and amplification of TLR genes should be employed. Since supplementation of the captive population is ongoing and samples are collected during routine health checks, TLR gene diversity can be monitored over time. These data can reveal how the genetic diversity at TLR genes changes over time and how management actions influence this diversity. Real-time monitoring of TLR gene diversity has the potential to add value to existing conservation genomic management actions.

Kakī

I was only able to identify four complete TLR genes within the kakī genome and was not able to characterise TLR gene diversity within the semi-wild population. A long-read reference genome is in the works for this species and will likely improve completeness and quality of the assembly and thus identification of full TLR sequences. It is also possible that the new PacBio

genome assembly will improve characterisation of TLR gene diversity within the kakī population. Once the new genome becomes available, this bioinformatic approach can be tested again to see whether it improves outcomes. If missingness in the resequencing data continues to cause difficulties for this future analysis, long-read resequencing of the population would likely improve characterisation of TLR gene diversity.

Due to the lack of TLR genes identified and characterised within kakī, I was not able to address the conservation management question of whether increasing founder representation also increases TLR gene diversity. Future analysis of the TLR gene diversity within the semi-wild kakī population will address this question through analysis of new and existing data.

Recommendations for Application of this Bioinformatic Approach

The following are recommendations for applying the bioinformatic approach developed in this thesis to other bird species. If the species reference genome is relatively incomplete, it may not be sufficient for identifying all TLR genes. Undertaking strategies to improve completeness of the genome may help, including alignment of multiple assemblies to fill in sequencing gaps. I also recommend undertaking deliberate sequencing strategies to maximise depth and minimise missingness of resequencing data. For example, if preliminary screening of resequencing data shows low depth and high missingness, it might be worth resequencing the libraries. Also, if there are individuals who are particularly important for analysis, then it may pay off to sequence them twice. As conservation genomic technologies move in the direction of long-reads both for resequencing data and reference genome assembly, they can readily be integrated into this approach and these longer sequences may facilitate identification and improve characterisation of TLR genes within species of interest.

One caveat for this method is that I conducted this Proof of Concept with three threatened bird species, so it is unknown whether it may be useful in other species with more complex genomes or heterozygous populations. Nevertheless, advancing technologies may provide

opportunities to apply this method widely to bird species with different levels of heterozygosity and genome complexity.

Horizon Scan for New Genomic Technologies

The landscape of conservation genomics technology is changing at a rapid pace (Kumar, Cowley, and Davis, 2019). Long-read sequencing is a recent development that will likely facilitate characterisation of functional diversity. In comparison to short-read sequences, which span hundreds of base pairs, long-reads span thousands of base pairs. Having longer reads limits gaps and errors to produce a more complete and high quality genome assembly than short-read sequencing (Pollard et al. 2018). A long-read genome may improve identification of genes within the reference genome, and long-read resequencing of individual genomes may improve alignment of population resequencing data to the reference genome. Also, long reads allow for greater assembly of highly repetitive regions which may enable identification of more complex and repetitive immune genes like MHC (Huddleston et al. 2014).

Another technology that is on the horizon is pangenomes, which combine reference genome sequences for several members of a population in order to better represent all genomic variation within a species (Tettelin et al. 2005; Wold et al. 2021, non-peer-reviewed pre-print). This approach may be particularly valuable for characterisation of genes that have to do with disease resistance and immune function (Bayer et al. 2020). Pangeomic resources would likely reveal previously undetected differentiation within TLR regions and may be especially important for populations with high heterozygosity.

Data Accessibility

Genomic data provided in this manuscript will be made available through a password protected server on the Conservation, Systematics and Evolution Research Team's website (http://www.ucconsert.org/data/). Tūturuatu, kākāriki karaka, and kakī are all taonga (treasured)

species. For Māori and Moriori (the Indigenous people of Aotearoa), all genomic data obtained from taonga species have whakapapa (genealogy that includes people, plants and animals, mountains, rivers and winds) and are therefore taonga in their own right. Thus, these data are tapu (sacred) and tikanga (customary practices, protocols, and ethics) determine how people interact with it. To this end, the passwords for the genomic data in this thesis will be made available to researchers on the recommendation of the kaitiaki (guardians) for the iwi and imi (tribes) that affiliate with tūturuatu, kākāriki karaka, and kakī.

References

Alcaide, M., and S. V. Edwards. 2011. "Molecular Evolution of the Toll-Like Receptor Multigene Family in Birds." *Molecular Biology and Evolution* 28 (5): 1703–15. https://doi.org/10.1093/molbev/msq351.

Alley, MR, and BD Gartrell. 2019. "Wildlife Diseases in New Zealand: Recent Findings and Future Challenges." *New Zealand Veterinary Journal* 67 (1): 1–11. https://doi.org/10.1080/00480169.2018.1520656.

Ballou, Jonathan D. 1993. "Assessing the Risks of Infectious Diseases in Captive Breeding and Reintroduction Programs." *Journal of Zoo and Wildlife Medicine* 24 (3): 327–35.

Barnett, Derek W., Erik K. Garrison, Aaron R. Quinlan, Michael P. Strömberg, and Gabor T. Marth. 2011. "BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files." *Bioinformatics* 27 (12): 1691–92. https://doi.org/10.1093/bioinformatics/btr174.

Barreiro, Luis B., Meriem Ben-Ali, Hélène Quach, Guillaume Laval, Etienne Patin, Joseph K. Pickrell, Christiane Bouchier, et al. 2009. "Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense." *PLOS Genetics* 5 (7): e1000562. https://doi.org/10.1371/journal.pgen.1000562.

Bayer, Philipp E., Agnieszka A. Golicz, Armin Scheben, Jacqueline Batley, and David Edwards. 2020. "Plant Pan-Genomes Are the New Reference." *Nature Plants* 6 (8): 914–20. https://doi.org/10.1038/s41477-020-0733-0.

Beutler, B., and M. Rehli. 2002. "Evolution of the TIR, Tolls and TLRs: Functional Inferences from Computational Biology." In *Toll-Like Receptor Family Members and Their Ligands*, edited by Bruce Beutler and Hermann Wagner, 1–21. Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-59430-4_1.

Bolger, AM, Lohse, M., and B Usadel. 2014. *Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. Bioinformatics, Btu170.* (version 0.35).

Bolte, Annette L., Josef Meurer, and Erhard F. Kaleta. 1999. "Avian Host Spectrum of Avipoxviruses." *Avian Pathology* 28 (5): 415–32. https://doi.org/10.1080/03079459994434.

Bonneaud, Camille, Susan L. Balenger, Jiangwen Zhang, Scott V. Edwards, and Geoffrey E. Hill. 2012. "Innate Immunity and the Evolution of Resistance to an Emerging Infectious Disease in a Wild Bird." *Molecular Ecology* 21 (11): 2628–39. https://doi.org/10.1111/j.1365-294X.2012.05551.x.

Boyle, D. B., and H. G. Heine. 1994. "Influence of Dose and Route of Inoculation on Responses of Chickens to Recombinant Fowlpox Virus Vaccines." *Veterinary Microbiology* 41 (1–2): 173–81. https://doi.org/10.1016/0378-1135(94)90146-5.

Boyle, David B. 2007. "Genus Avipoxvirus." *Poxviruses*, 217–51. https://doi.org/10.1007/978-3-7643-7557-7_11.

Brady, G, and AG Bowie. 2014. "Innate Immune Activation of NFκB and Its Antagonism by Poxviruses." Cytokine & Growth Factor Reviews. Cytokine Growth Factor Rev. October 2014. https://doi.org/10.1016/j.cytogfr.2014.07.004.

Brandies, Parice A., Belinda R. Wright, Carolyn J. Hogg, Catherine E. Grueber, and Katherine Belov. 2020. "Characterisation of Reproductive Gene Diversity in the Endangered Tasmanian Devil." *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13295.

Brucker, Robert M., Lisa J. Funkhouser, Shefali Setia, Rini Pauly, and Seth R. Bordenstein. 2012. "Insect Innate Immunity Database (IIID): An Annotation Tool for Identifying Immune Genes in Insect Genomes." *PLOS ONE* 7 (9): e45125. https://doi.org/10.1371/journal.pone.0045125.

Cao, Hua, Peihong Dai, Weiyi Wang, Hao Li, Jianda Yuan, Fangjin Wang, Chee-Mun Fang, et al. 2012. “Innate Immune Response of Human Plasmacytoid Dendritic Cells to Poxvirus Infection Is Subverted by Vaccinia E3 via Its Z-DNA/RNA Binding Domain.” *PLOS ONE* 7 (5): e36823. https://doi.org/10.1371/journal.pone.0036823.

Chávez-Treviño, Alejandro, Ricardo Canales-del-Castillo, Irene Ruvalcaba-Ortega, Diana Reséndez-Pérez, José Ignacio González-Rojas, Antonio Guzmán-Velasco, and Arvind O. Panjabi. 2017. “Primer Development for Amplification of Toll-like Genes for the Assessment of Adaptive Genetic Diversity in Vulnerable Grassland Bird Species.” *Conservation Genetics Resources* 9 (3): 385–87. https://doi.org/10.1007/s12686-017-0690-8.

Cheng, Zhimin, Mengxuan Li, Yao Wang, Tongjie Chai, Yumei Cai, and Ning Li. 2020. “Pathogenicity and Immune Responses of Aspergillus Fumigatus Infection in Chickens.” *Frontiers in Veterinary Science* 7. https://doi.org/10.3389/fvets.2020.00143.

Clark, Rachael, and Thomas Kupper. 2005. “Old Meets New: The Interaction Between Innate and Adaptive Immunity.” *Journal of Investigative Dermatology* 125 (4): 629–37. https://doi.org/10.1111/j.0022-202X.2005.23856.x.

Copeland, Claudia S., Manja Marz, Dominic Rose, Jana Hertel, Paul J. Brindley, Clara Bermudez Santana, Stephanie Kehr, Camille Stephan-Otto Attolini, and Peter F. Stadler. 2009. “Homology-Based Annotation of Non-Coding RNAs in the Genomes of Schistosoma Mansoni and Schistosoma Japonicum.” *BMC Genomics* 10 (1): 464. https://doi.org/10.1186/1471-2164-10-464.

Cui, Jian, Yuanyuan Cheng, and Katherine Belov. 2015. “Diversity in the Toll-like Receptor Genes of the Tasmanian Devil (Sarcophilus Harrisii).” *Immunogenetics* 67 (3): 195–201. https://doi.org/10.1007/s00251-014-0823-0.

Dalton, Desiré L., Elaine Vermaak, Hanneline A. Smit-Robinson, and Antoinette Kotze. 2016. “Lack of Diversity at Innate Immunity Toll-like Receptor Genes in the Critically Endangered White-Winged Flufftail (Sarothrura Ayresi).” *Scientific Reports* 6 (1): 1–8. https://doi.org/10.1038/srep36757.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Davies, Charli, Martin Taylor, Martijn Hammers, Terry Burke, J Komdeur, Hannah Dugdale, and David Richardson. 2021. “Contemporary Evolution of the Viral-Sensing TLR3 Gene in an Isolated Vertebrate Population.” Preprint. Preprints. https://doi.org/10.22541/au.160992924.47833640/v1.

De-Kayne, Rishi, David Frei, Ryan Greenway, Sofia L. Mendes, Cas Retel, and Philine G. D. Feulner. 2020 “Sequencing Platform Shifts Provide Opportunities but Pose Challenges for Combining Genomic Data Sets.” *Molecular Ecology Resources*. Accessed December 29, 2020. https://doi.org/10.1111/1755-0998.13309.

Dowding, John E. 2013. “Reducing the Risk of Extinction of a Globally Threatened Shorebird: Translocations of the Shore Plover (Thinornis Novaeseelandiae), 1990-2012,” 15.

Dowding, John E, and Elaine C Murphy. 2001. “The Impact of Predation by Introduced Mammals on Endemic Shorebirds in New Zealand: A Conservation Perspective.” *Biological Conservation* 99 (1): 47–64. https://doi.org/10.1016/S0006-3207(00)00187-7.

Doyle, R.M., R. Perez-Enriquez, M. Takagi, and N. Taniguchi. 2001. “Selective Recovery of Founder Genetic Diversity in Aquacultural Broodstocks and Captive, Endangered Fish Populations.” *Genetica* 111 (1): 291–304. https://doi.org/10.1023/A:1013772205330.

Ewen JG, Armstrong DP, Parker KA, Seddon PJ. 2012. *Reintroduction Biology: Integrating Science and Management*. Conservation Science and Practice Series. Oxford, UK: Wiley-Blackwell.

Fellah, Jaffredo, and Dunon. 2008. “Development of the Avian Immune System.” In *Avian*

Immunology, 51–66. 4.
Feng, S, Stiller, J, and Deng, Y. 2020. "Dense Sampling of Bird Diversity Increases Power of Comparative Genomics." *Nature*, 587, , 252–57. https://doi.org/ttps://doi.org/10.1038/s41586-020-2873-9.
Feuillet, Vincent, Samir Medjane, Isabelle Mondor, Olivier Demaria, Philippe P. Pagni, Jorge E. Galán, Richard A. Flavell, and Lena Alexopoulou. 2006. "Involvement of Toll-like Receptor 5 in the Recognition of Flagellated Bacteria." *Proceedings of the National Academy of Sciences* 103 (33): 12487–92. https://doi.org/10.1073/pnas.0605200103.
Forsdick, Natalie. 2020. "Assessment of the Impacts of Anthropogenic Hybridisation in a Threatened Non-Model Bird Species through the Development of Genomic Resources with Implications for Conservation." University of Otago.
Galla, Stepanie J. 2019. "Conservation Genomic Management of Two Critically Endangered New Zealand Birds." University of Canterbury School of Biological Science.
Galla, Stephanie J., Roger Moraga, Liz Brown, Simone Cleland, Marc P. Hoeppner, Richard F. Maloney, Anne Richardson, Lyndon Slater, Anna W. Santure, and Tammy E. Steeves. 2020. "A Comparison of Pedigree, Genetic and Genomic Estimates of Relatedness for Informing Pairing Decisions in Two Critically Endangered Birds: Implications for Conservation Breeding Programmes Worldwide." *Evolutionary Applications*. https://doi.org/10.1111/eva.12916.
Galla, Stephanie, and Tammy Steeves. 2018. "'Orange-Fronted Kākāriki Blumine Pedigree Report, July 2018.'"
Gatrell, BD, MR Alley, J Pauli, R Collen, and K Barlow. 2002. "An Outbreak of Poxvirus Infection in Captive Shore Plovers." *Kokako* 9 (2): 9–11.
Giotis, Efstathios S., Stephen M. Laidlaw, Susanna R. Bidgood, David Albrecht, Jemima J. Burden, Rebecca C. Robey, Jason Mercer, and Michael A. Skinner. 2020. "Modulation of Early Host Innate Immune Response by an Avipox Vaccine Virus' Lateral Body Protein." *Biomedicines* 8 (12): 634. https://doi.org/10.3390/biomedicines8120634.
Glassock, Gael L., Catherine E. Grueber, Katherine Belov, and Carolyn J. Hogg. 2021. "Reducing the Extinction Risk of Populations Threatened by Infectious Diseases." *Diversity* 13 (2): 63. https://doi.org/10.3390/d13020063.
Goltsman, Eugene, Isaac Ho, and Daniel Rokhsar. 2017. "Meraculous-2D: Haplotype-Sensitive Assembly of Highly Heterozygous Genomes." *ArXiv:1703.09852 [q-Bio]*, March. http://arxiv.org/abs/1703.09852.
Grueber, Catherine E. 2015. "Comparative Genomics for Biodiversity Conservation." *Computational and Structural Biotechnology Journal* 13 (January): 370–75. https://doi.org/10.1016/j.csbj.2015.05.003.
Grueber, Catherine E., Gabrielle J. Knafler, Tania M. King, Alistair M. Senior, Stefanie Grosser, Bruce Robertson, Kerry A. Weston, Patricia Brekke, Christian L. W. Harris, and Ian G. Jamieson. 2015. "Toll-like Receptor Diversity in 10 Threatened Bird Species: Relationship with Microsatellite Heterozygosity." *S* 16 (3): 595–611. https://doi.org/10.1007/s10592-014-0685-x.
Grueber, Catherine E., Jolene T. Sutton, Sol Heber, James V. Briskie, Ian G. Jamieson, and Bruce C. Robertson. 2017. "Reciprocal Translocation of Small Numbers of Inbred Individuals Rescues Immunogenetic Diversity." *Molecular Ecology* 26 (10): 2660–73. https://doi.org/10.1111/mec.14063.
Grueber, Catherine E., Graham P. Wallis, and Ian G. Jamieson. 2014. "Episodic Positive Selection in the Evolution of Avian Toll-Like Receptor Innate Immunity Genes." Edited by Hans Ellegren. *PLoS ONE* 9 (3): e89632. https://doi.org/10.1371/journal.pone.0089632.
Grueber, Catherine E., Graham P. Wallis, Tania M. King, and Ian G. Jamieson. 2012. "Variation at Innate Immunity Toll-Like Receptor Genes in a Bottlenecked Population of a New Zealand Robin." Edited by Hans Ellegren. *PLoS ONE* 7 (9): e45011.

https://doi.org/10.1371/journal.pone.0045011.
Ha HJ, Alley MR, Cahill JI, Howe L, and Gartrell BD. 2009. "The Prevalence of Psittacine Beak and Feather Disease Virus Infection in Native Parrots in New Zealand." *New Zealand Veterinary Journal* 57 (1): 50–52. https://doi.org/10.1080/00480169.2009.36868.
Ha, Hye Jeong, Maurice Alley, Laryssa Howe, and Brett Gartrell. 2013. "Evaluation of the Pathogenicity of Avipoxvirus Strains Isolated from Wild Birds in New Zealand and the Efficacy of a Fowlpox Vaccine in Passerines." *Veterinary Microbiology* 165 (3–4): 268–74. https://doi.org/10.1016/j.vetmic.2013.03.015.
Ha, Hye Jeong, Laryssa Howe, Maurice Alley, and Brett Gartrell. 2011. "The Phylogenetic Analysis of Avipoxvirus in New Zealand." *Veterinary Microbiology* 150 (1): 80–87. https://doi.org/10.1016/j.vetmic.2011.01.011.
Hansen, W. 1999. "Avian Pox." Report 1999–0001. Information and Technology Report. Reston, VA. USGS Publications Warehouse. http://pubs.er.usgs.gov/publication/2001161.
Harte, Mary T., Ismar R. Haga, Geraldine Maloney, Pearl Gray, Patrick C. Reading, Nathan W. Bartlett, Geoffrey L. Smith, Andrew Bowie, and Luke A.J. O'Neill. 2003. "The Poxvirus Protein A52R Targets Toll-like Receptor Signaling Complexes to Suppress Host Defense." *Journal of Experimental Medicine* 197 (3): 343–51. https://doi.org/10.1084/jem.20021652.
Hartmann, Stefanie A., H. Martin Schaefer, and Gernot Segelbacher. 2014. "Genetic Depletion at Adaptive but Not Neutral Loci in an Endangered Bird Species." *Molecular Ecology* 23 (23): 5712–25. https://doi.org/10.1111/mec.12975.
Hendaus, Mohamed A, Fatima A Jomha, and Ahmed H Alhammadi. 2015. "Virus-Induced Secondary Bacterial Infection: A Concise Review." *Therapeutics and Clinical Risk Management* 11 (August): 1265–71. https://doi.org/10.2147/TCRM.S87789.
Heng, Jianfu, Jianguo Su, Teng Huang, Jie Dong, and Lijun Chen. 2011. "The Polymorphism and Haplotype of TLR3 Gene in Grass Carp (Ctenopharyngodon Idella) and Their Associations with Susceptibility/Resistance to Grass Carp Reovirus." *Fish & Shellfish Immunology* 30 (1): 45–50. https://doi.org/10.1016/j.fsi.2010.09.004.
Huddleston, John, Swati Ranade, Maika Malig, Francesca Antonacci, Mark Chaisson, Lawrence Hon, Peter H. Sudmant, et al. 2014. "Reconstructing Complex Regions of Genomes Using Long-Read Sequencing Technology." *Genome Research* 24 (4): 688–96. https://doi.org/10.1101/gr.168450.113.
Irizarry, Kristopher J. L., Doug Bryant, Jordan Kalish, Curtis Eng, Peggy L. Schmidt, Gini Barrett, and Margaret C. Barr. 2016. "Integrating Genomic Data Sets for Knowledge Discovery: An Informed Approach to Management of Captive Endangered Species." *International Journal of Genomics* 2016: e2374610. https://doi.org/10.1155/2016/2374610.
Jackson, R, RS Morris, and W Boardman. 2000. "Development of a Method for Evaluating the Risk to New Zealand's Indigenous Fauna from the Introduction of Exotic Diseases and Pests - Including a Case Study on Native Parrots," 93.
Jiang, Ping-ping, Qiu-lei Lang, Sheng-guo Fang, Ping Ding, and Li-ming Chen. 2005. "A Genetic Diversity Comparison between Captive Individuals and Wild Individuals of Elliot's Pheasant (Syrmaticus Ellioti) Using Mitochondrial DNA." *Journal of Zhejiang University. Science. B* 6 (5): 413–17. https://doi.org/10.1631/jzus.2005.B0413.
Johnston, J. B., and Grant McFadden. 2003. "Poxvirus Immunomodulatory Strategies: Current Perspectives." *Journal of Virology* 77 (11): 6093–6100. https://doi.org/10.1128/JVI.77.11.6093-6100.2003.
Kannaki, T.R., M.R. Reddy, M. Shanmugam, P.C. Verma, and R.P. Sharma. 2010. "Chicken Toll-like Receptors and Their Role in Immunity." *World's Poultry Science Journal* 66 (4): 727–38. https://doi.org/10.1017/S0043933910000693.
Kawasaki, Takumi, and Taro Kawai. 2014. "Toll-Like Receptor Signaling Pathways." *Frontiers in*

Immunology 5. https://doi.org/10.3389/fimmu.2014.00461.
Kearvell, J.C., and A.J. Legault. 2017. "Orange-Fronted Parakeet." 2017. www.nzbirdsonline.org.nz.
Kearvell, Jonathan C, and Andrew D Grant. 2003. "The Orange-Fronted Parakeet (Cyanoramphusmalherbi) Is a Distinct Species: A Review of Recent Research into Its Taxonomy and Systematic Relationship within the Genus Cyanoramphus," 10.
Kearvell, Jonathan C, James R Young, and Andrew D Grant. 2002. "Comparative Ecology of Sympatric Orange-Fronted Parakeets (Cyanoramphus Malherbi) and Yellow-Crowned Parakeets (C. Auriceps), South Island, New Zealand." *NEW ZEALAND JOURNAL OF ECOLOGY* 26 (2): 10.
Keestra, A. Marijke, Marcel R. de Zoete, Rémon A. M. H. van Aubel, and Jos P. M. van Putten. 2008. "Functional Characterization of Chicken TLR5 Reveals Species-Specific Recognition of Flagellin." *Molecular Immunology* 45 (5): 1298–1307. https://doi.org/10.1016/j.molimm.2007.09.013.
Kennedy, Richard B., Gregory A. Poland, Inna G. Ovsyannikova, Ann L. Oberg, Yan W. Asmann, Diane E. Grill, Robert A. Vierkant, and Robert M. Jacobson. 2016. "Impaired Innate, Humoral, and Cellular Immunity despite a Take in Smallpox Vaccine Recipients." *Vaccine* 34 (28): 3283–90. https://doi.org/10.1016/j.vaccine.2016.05.005.
Kielbasa, S. M., R. Wan, K. Sato, P. Horton, and M. C. Frith. 2011. "Adaptive Seeds Tame Genomic Sequence Comparison." *Genome Research* 21 (3): 487–93. https://doi.org/10.1101/gr.113985.110.
King, C, M Stone, and J Wang. 2003. "An Outbreak of Disease Caused by Psittacinepoxvirus in Rosellas." *Surveillance* 30 (3): 11–13.
Knafler, Gabrielle, Catherine Grueber, Jolene Sutton, and Ian Jamieson. 2017. "Differential Patterns of Diversity at Microsatellite, MHC, and TLR Loci in Bottlenecked South Island Saddleback Populations." *New Zealand Journal of Ecology* 41 (1). https://doi.org/10.20417/nzjecol.41.8.
Knafler, Gabrielle J., Luis Ortiz-catedral, Bethany Jackson, Arvind Varsani, Catherine E. Grueber, Bruce C. Robertson, and Ian G. Jamieson. 2016. "Comparison of Beak and Feather Disease Virus Prevalence and Immunity-Associated Genetic Diversity over Time in an Island Population of Red-Crowned Parakeets." *Archives of Virology; Wien* 161 (4): 811–20. http://dx.doi.org.ezproxy.canterbury.ac.nz/10.1007/s00705-015-2717-3.
Kreuger, Felix. 2020. *Trim Galore* (version 0.6.5). Retrieved from https://github.com/FelixKrueger/TrimGalore.
Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin. 2011. "Computational Methods for Gene Orthology Inference." *Briefings in Bioinformatics* 12 (5): 379–91. https://doi.org/10.1093/bib/bbr030.
Kumar, Kishore R., Mark J. Cowley, and Ryan L. Davis. 2019. "Next-Generation Sequencing and Emerging Technologies." *Seminars in Thrombosis and Hemostasis* 45 (7): 661–73. https://doi.org/10.1055/s-0039-1688446.
Lacy, R.C., J.D. Ballou, and J.P. Pollak. 2012. "PMx: Software Package for Demographic and Genetic Analysis and Management of Pedigreed Populations." *Methods in Ecology & Evolution* 3: 433–37.
Lacy, Robert C. 1987. "Loss of Genetic Diversity from Managed Populations: Interacting Effects of Drift, Mutation, Immigration, Selection, and Population Subdivision." *Conservation Biology* 1 (2): 143–58.
Leishangthem, G D, N D Singh, R S Brar, and H S Banga. 2015. "Aspergillosis in Avian Species: A Review" 3 (1): 14.
Lester, Sandra N., and Kui Li. 2014. "Toll-Like Receptors in Antiviral Innate Immunity." *Journal of Molecular Biology* 426 (6): 1246–64. https://doi.org/10.1016/j.jmb.2013.11.024.
Letunic, Ivica, Supriya Khedkar, and Peer Bork. 2021. "SMART: Recent Updates, New

Developments and Status in 2020." *Nucleic Acids Research* 49 (D1): D458–60. https://doi.org/10.1093/nar/gkaa937.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.
Liu, Ting, Lingyun Zhang, Donghyun Joo, and Shao-Cong Sun. 2017. "NF-KB Signaling in Inflammation." *Signal Transduction and Targeted Therapy* 2 (1): 1–9. https://doi.org/10.1038/sigtrans.2017.23.
Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. "SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler." *GigaScience* 1 (2047-217X-1–18). https://doi.org/10.1186/2047-217X-1-18.
Madden, Thomas. 2013. *The BLAST Sequence Analysis Tool. The NCBI Handbook [Internet]. 2nd Edition*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK153387/.
Mander C, Gartrell B, and McInnes K. 2003. "Psittacine Beak and Feather Disease in a Wild Eastern Rosella." *NZVA Companion Animal Quarterly* 14 (4): 24–25.
Marsden, C. D., H. Verberkmoes, R. Thomas, R. K. Wayne, and B. K. Mable. 2013. "Pedigrees, MHC and Microsatellites: An Integrated Approach for Genetic Management of Captive African Wild Dogs (Lycaon Pictus)." *Conservation Genetics* 14 (1): 171–83. https://doi.org/10.1007/s10592-012-0440-0.
Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads," 3.
Matsushima, Norio, Takanori Tanaka, Purevjav Enkhbayar, Tomoko Mikami, Masae Taga, Keiko Yamada, and Yoshio Kuroki. 2007. "Comparative Sequence Analysis of Leucine-Rich Repeats (LRRs) within Vertebrate Toll-like Receptors." *BMC Genomics* 8 (1): 124. https://doi.org/10.1186/1471-2164-8-124.
McLennan, E. A., C. E. Grueber, P. Wise, K. Belov, and C. J. Hogg. 2020. "Mixing Genetically Differentiated Populations Successfully Boosts Diversity of an Endangered Carnivore." *Animal Conservation* 23 (6): 700–712. https://doi.org/10.1111/acv.12589.
Medzhitov, Ruslan. 2001. "Toll-like Receptors and Innate Immunity." *Nature Reviews Immunology* 1 (2): 135–45. https://doi.org/10.1038/35100529.
Minias, Piotr, Ewa Pikus, Linda A Whittingham, and Peter O Dunn. 2019. "Evolution of Copy Number at the MHC Varies across the Avian Tree of Life." Edited by Yoshihito Niimura. *Genome Biology and Evolution* 11 (1): 17–28. https://doi.org/10.1093/gbe/evy253.
Moraga, Roger. 2018. *SubSampler_SNPcaller*. Tea Break Bioinformatics. https://github.com/Lanilen/SubSampler_SNPcaller.
Morris, Katrina M., Belinda Wright, Catherine E. Grueber, Carolyn Hogg, and Katherine Belov. 2015. "Lack of Genetic Diversity across Diverse Immune Genes in an Endangered Mammal, the Tasmanian Devil (*Sarcophilus Harrisii*)." *Molecular Ecology* 24 (15): 3860–72. https://doi.org/10.1111/mec.13291.
Morrison, Caitlin E., Carolyn J. Hogg, Rosemary Gales, Rebecca N. Johnson, and Catherine E. Grueber. 2020. "Low Innate Immune-Gene Diversity in the Critically Endangered Orange-Bellied Parrot (*Neophema Chrysogaster*)." *Emu - Austral Ornithology* 120 (1): 56–64. https://doi.org/10.1080/01584197.2019.1686994.
Mueller, Ralf C., Nicolai Mallig, Jacqueline Smith, Lél Eöery, Richard I. Kuo, and Robert H. S. Kraus. 2020. "Avian Immunome DB: An Example of a User-Friendly Interface for Extracting Genetic Information." *BMC Bioinformatics* 21 (1): 502. https://doi.org/10.1186/s12859-020-03764-3.
Narayanan, Kannan Badri, and Hyun Ho Park. 2015. "Toll/Interleukin-1 Receptor (TIR) Domain-

Mediated Cellular Signaling Pathways.” *Apoptosis* 20 (2): 196–209. https://doi.org/10.1007/s10495-014-1073-1.
NCBI. 2020. *BLAST Results: Expect Values, Part 1*. https://www.youtube.com/watch?v=ZN3RrXAe0uM.
Oda, Shun-ichiro, Edward Franklin, and Amir R. Khan. 2011. “Poxvirus A46 Protein Binds to TIR Domain-Containing Mal/TIRAP via an α-Helical Sub-Domain.” *Molecular Immunology* 48 (15–16): 2144–50. https://doi.org/10.1016/j.molimm.2011.07.014.
Offord, V., T. J. Coffey, and D. Werling. 2010. “LRRfinder: A Web Application for the Identification of Leucine-Rich Repeats and an Integrative Toll-like Receptor Database.” *Developmental & Comparative Immunology* 34 (10): 1035–41. https://doi.org/10.1016/j.dci.2010.05.004.
Overbeek, Ashley. 2020. “Pedigree Validation Using Genetic Markers in an Intensively-Managed Taonga Species, the Critically Endangered Kakī,” 8.
Palacios, Maria G., Joan E. Cunnick, David Vleck, and Carol M. Vleck. 2009. “Ontogeny of Innate and Adaptive Immune Defense Components in Free-Living Tree Swallows, Tachycineta Bicolor.” *Developmental & Comparative Immunology* 33 (4): 456–63. https://doi.org/10.1016/j.dci.2008.09.006.
Pasare, Chandrashekhar, and Ruslan Medzhitov. 2005. “Toll-Like Receptors: Linking Innate and Adaptive Immunity.” In *Mechanisms of Lymphocyte Activation and Immune Regulation X*, edited by Sudhir Gupta, William E. Paul, and Ralph Steinman, 11–18. Advances in Experimental Medicine and Biology. Boston, MA: Springer US. https://doi.org/10.1007/0-387-24180-9_2.
Pierce, RJ. 1984. “The Changed Distribution of Stilts in New Zealand.”
Pollard, Martin O, Deepti Gurdasani, Alexander J Mentzer, Tarryn Porter, and Manjinder S Sandhu. 2018. “Long Reads: Their Purpose and Place.” *Human Molecular Genetics* 27 (R2): R234–41. https://doi.org/10.1093/hmg/ddy177.
Pulendran, Bali, and Rafi Ahmed. 2006. “Translating Innate Immunity into Immunological Memory: Implications for Vaccine Development.” *Cell* 124 (4): 849–63. https://doi.org/10.1016/j.cell.2006.02.019.
“QC Fail Sequencing » Illumina 2 Colour Chemistry Can Overcall High Confidence G Bases.” 2016. Accessed December 16, 2020. https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases/.
Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.
Raven, Nynke, Simeon Lisovski, Marcel Klaassen, Nathan Lo, Thomas Madsen, Simon Y. W. Ho, and Beata Ujvari. 2017. “Purifying Selection and Concerted Evolution of RNA-Sensing Toll-like Receptors in Migratory Waders.” *Infection, Genetics and Evolution* 53 (September): 135–45. https://doi.org/10.1016/j.meegid.2017.05.012.
Rawlence, N. J., M. Kennedy, R. P. Scofield, A. J. D. Tennyson, P. Boussès, and G. K. Chambers. 2015. “Taxonomic Status of Cyanoramphus Parakeets on the Auckland Islands and Implications for the Validity of the Orange-Fronted Parakeet as a Discrete Species.” *Journal of the Royal Society of New Zealand* 45 (4): 197–211. https://doi.org/10.1080/03036758.2015.1084337.
Reed, Christine Elva Margaret, New Zealand, and Biodiversity Recovery Unit. 1998. *Management Plan for Captive Black Stilts*. Wellington, N.Z.: Biodiversity Recovery Unit, Dept. of Conservation.
Riper, Charles van, and Donald J. Forrester. 2007. “Avian Pox.” In *Infectious Diseases of Wild Birds*, 131–76. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470344668.ch6.
Romanutti, Carina, Leticia Keller, and Flavia Adriana Zanetti. 2020. “Current Status of Virus-Vectored Vaccines against Pathogens That Affect Poultry.” *Vaccine* 38 (45): 6990–7001.

https://doi.org/10.1016/j.vaccine.2020.09.013.
Rozas, Julio. “DnaSP v6. Documentation,” 112.
Samuelsson, Christofer, Jürgen Hausmann, Henning Lauterbach, Michaela Schmidt, Shizuo Akira, Hermann Wagner, Paul Chaplin, Mark Suter, Meredith O’Keeffe, and Hubertus Hochrein. 2008. “Survival of Lethal Poxvirus Infection in Mice Depends on TLR9, and Therapeutic Vaccination Provides Protection.” *Journal of Clinical Investigation* 118 (5): 1776–84. https://doi.org/10.1172/JCI33940.
Sauna, Zuben E., and Chava Kimchi-Sarfaty. 2011. “Understanding the Contribution of Synonymous Mutations to Human Disease.” *Nature Reviews Genetics* 12 (10): 683–91. https://doi.org/10.1038/nrg3051.
Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting. 1998. “SMART, a Simple Modular Architecture Research Tool: Identification of Signaling Domains.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (11): 5857–64. https://doi.org/10.1073/pnas.95.11.5857.
Sedlazeck, Fritz J., Philipp Rescheneder, and Arndt von Haeseler. 2013. “NextGenMap: Fast and Accurate Read Mapping in Highly Polymorphic Genomes.” *Bioinformatics* 29 (21): 2790–91. https://doi.org/10.1093/bioinformatics/btt468.
Seet, Bruce T., J. B. Johnston, Craig R. Brunetti, John W. Barrett, Helen Everett, Cheryl Cameron, Joanna Sypula, Steven H. Nazarian, Alexandra Lucas, and Grant McFadden. 2003. “Poxviruses and Immune Evasion.” *Annual Review of Immunology* 21: 377–423. https://doi.org/10.1146/annurev.immunol.21.120601.141049.
Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.
Singh, Bhanu P., R. S. Chauhan, and Lokesh K. Singhal. 2003. “Toll-like Receptors and Their Role in Innate Immunity.” *Current Science* Vol. 85 (No. 8): 1156–64.
Smeds, Linnéa, and Axel Künstner. 2011. “ConDeTri - A Content Dependent Read Trimmer for Illumina Data.” *PloS One* 6 (October): e26314. https://doi.org/10.1371/journal.pone.0026314.
Sommer, Simone. 2005. “The Importance of Immune Gene Variability (MHC) in Evolutionary Ecology and Conservation.” *Frontiers in Zoology* 2 (1): 16. https://doi.org/10.1186/1742-9994-2-16.
Stack, Julianne, Ismar R. Haga, Martina Schröder, Nathan W. Bartlett, Geraldine Maloney, Patrick C. Reading, Katherine A. Fitzgerald, Geoffrey L. Smith, and Andrew G. Bowie. 2005. “Vaccinia Virus Protein A46R Targets Multiple Toll-like–Interleukin-1 Receptor Adaptors and Contributes to Virulence.” *Journal of Experimental Medicine* 201 (6): 1007–18. https://doi.org/10.1084/jem.20041442.
Stanford, Marianne M., Grant McFadden, Gunasegaran Karupiah, and Geeta Chaudhri. 2007. “Immunopathogenesis of Poxvirus Infections: Forecasting the Impending Storm.” *Immunology and Cell Biology* 85 (2): 93–102. https://doi.org/10.1038/sj.icb.7100033.
Steeves, Tammy E., Richard F. Maloney, Marie L. Hale, Jason M. Tylianakis, and Neil J. Gemmell. 2010. “Genetic Analyses Reveal Hybridization but No Hybrid Swarm in One of the World’s Rarest Birds.” *Molecular Ecology* 19 (23): 5090–5100. https://doi.org/10.1111/j.1365-294X.2010.04895.x.
Tajima, F. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123 (3): 585–95.
Takeda, Kiyoshi, and Shizuo Akira. 2005. “Toll-like Receptors in Innate Immunity.” *International Immunology* 17 (1): 1–14. https://doi.org/10.1093/intimm/dxh186.
Temperley, Nicholas D., Sofia Berlin, Ian R. Paton, Darren K. Griffin, and David W. Burt. 2008. “Evolution of the Chicken Toll-like Receptor Gene Family: A Story of Gene Gain and

Gene Loss." *BMC Genomics* 9 (February): 62. https://doi.org/10.1186/1471-2164-9-62.
Thompson, Joseph M., and Akiko Iwasaki. 2008. "TOLL-LIKE RECEPTORS REGULATION OF VIRAL INFECTION AND DISEASE." *Advanced Drug Delivery Reviews* 60 (7): 786–94. https://doi.org/10.1016/j.addr.2007.11.003.
Tompkins, D. M., and D. M. Gleeson. 2006. "Relationship between Avian Malaria Distribution and an Exotic Invasive Mosquito in New Zealand." *Journal of the Royal Society of New Zealand* 36 (2): 51–62. https://doi.org/10.1080/03014223.2006.9517799.
Tompkins, Daniel M., Scott Carver, Menna E. Jones, Martin Krkošek, and Lee F. Skerratt. 2015. "Emerging Infectious Diseases of Wildlife: A Critical Perspective." *Trends in Parasitology* 31 (4): 149–59. https://doi.org/10.1016/j.pt.2015.01.007.
Velová, Hana, Maria W. Gutowska-Ding, David W. Burt, and Michal Vinkler. 2018. "Toll-Like Receptor Evolution in Birds: Gene Duplication, Pseudogenization, and Diversifying Selection." *Molecular Biology and Evolution* 35 (9): 2170–84. https://doi.org/10.1093/molbev/msy119.
Weber, Michel J. 2005. "New Human and Mouse MicroRNA Genes Found by Homology Search." *The FEBS Journal* 272 (1): 59–73. https://doi.org/10.1111/j.1432-1033.2004.04389.x.
Weli, Simon C., and Morten Tryland. 2011. "Avipoxviruses: Infection Biology and Their Use as Vaccine Vectors." *Virology Journal* 8 (1): 49. https://doi.org/10.1186/1743-422X-8-49.
Wheeler, David L., Deanna M. Church, Scott Federhen, Alex E. Lash, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, et al. 2003. "Database Resources of the National Center for Biotechnology." *Nucleic Acids Research* 31 (1): 28–33.
Whitehorn, Penelope R., Matthew C. Tinsley, Mark J. F. Brown, Ben Darvill, and Dave Goulson. 2011. "Genetic Diversity, Parasite Prevalence and Immunity in Wild Bumblebees." *Proceedings of the Royal Society B: Biological Sciences* 278 (1709): 1195–1202. https://doi.org/10.1098/rspb.2010.1550.
Yilmaz, Ahmet, Shixue Shen, David L. Adelson, Suresh Xavier, and James J. Zhu. 2005. "Identification and Sequence Analysis of Chicken Toll-like Receptors." *Immunogenetics* 56 (10): 743–53. https://doi.org/10.1007/s00251-004-0740-8.
Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, et al. 2014. "Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation." *Science* 346 (6215): 1311–20. https://doi.org/10.1126/science.1251385.
Zhu, Ying, Catherine Grueber, Yudong Li, Ming He, Lan Hu, Ke He, Honhyi Liu, Hemin Zhang, and Honglin Wu. 2020. "MHC-Associated Baylisascaris Schroederi Load Informs the Giant Panda Reintroduction Program." *International Journal for Parasitology: Parasites and Wildlife*, May. https://doi.org/10.1016/j.ijppaw.2020.05.010.
Zimin, AZ, Marçais C, Puiu D, Roberts D, Salzberg SL, and Yorke JA. 2013. "The MaSuRCA Genome Assembler." *Bioinformatics* 29 (21): 2669–77.

Appendices

Contribution Statement

The genomic resources that I used for my thesis were generated by other members of our research team for aligned projects. Specifically, Roger Moraga at Tea Break Bioinformatics assembled the tūturuatu and kākāriki karaka genomes, and Nat Forsdick assembled the kakī genome. Stephanie Galla extracted the kākāriki karaka and kakī resequencing samples, and Ilina Cubrinovska extracted the tūturuatu resequencing samples.

Many of the following scripts are a result of collaboration with and adaptation of code written by other members of our research group. I adapted the alignment and variant calling script from code written by Natalie Forsdick and Levi Collier-Robinson. I adapted the script for extracting unmapped reads from a script written by Jana Wold based on an analysis by Laine et al 2019. I also adapted the script for quantifying genome-wide read depth from a script written by Jana Wold. Guidance on all bioinformatic methods and code was provided by Roger Moraga.

Appendix A: Samples Used in Resequencing Analysis

Table A.1: Stud book identification number (SB#) for all tūturuatu, kākāriki karaka, and kakī samples used in this thesis. Parentage of samples is also reported for kākāriki karaka (see Chapter 3 for details). Abbreviations: C=captive, W=wild, T=translocated.

Tūturuatu	Kākāriki karaka		Kakī
SB #	SB #	Parentage	SB #
889	398	CxC	361
600	120	CxC	262
613	151	WxW	260
790	157	CxC	258
707	196	WxW	112

849	294	CxC	1791
852	386	CxC	1792
614	447	CxC	1848
723	450	WxW	1850
714	451	WxW	1986
649	452	WxW	2054
430	453	WxW	2058
856	487	CxC	2385
428	503	CxC	2310
125	539	CxC	2368
452	542	CxC	2356
847	543	CxC	2449
660	544	CxC	2446
848	545	CxC	2437
728	546	CxC	2497
903	582	CxC	2396
935	593	CxC	2420
867	595	CxC	2424
922	607	CxC	15_114
990	608	CxC	2252
991	612	CxC	2278
1030	614	CxC	2379
1010	615	CxC	2338
1002	616	CxC	2339
1003	619	CxC	2440
1025	623	CxC	2436
745	624	CxC	2512
1042	625	CxC	2516
1043	630	CxC	2517
1045	651	CxC	2476
1051	655	CxC	

1052	585	WxW
1075	584	WxW
1035	631	CxC
	686	TxT
	695	WxC
	696	WxC
	697	WxC
	698	WxC
	699	WxC
	700	TxT
	701	TxT
	703	TxT

Appendix B: Bash scripts associated with TLR identification and characterisation

BLAST alignment

The following lines of code were used to create a BLAST database for each species' reference genome and to BLAST reference TLRs against the database:

```
#make blast database for genome
makeblastdb -dbtype nucl species_genome.fasta
#blast reference TLR against genome database
blastn or tblastn -query reference_sequence.fasta -db reference genome -out out_file.txt
```

Alignment and Variant Calling Script

The following script was used for all three species alignments, with variation to file locations.

```
#!/bin/bash -e

ref=/data/Tuturuatu/ref_genomes/Maui_masurca.fa #Reference genome for alignment

rawdata=/data/Tuturuatu/data/raw/

datadir=/data/Tuturuatu/data/trimmed_galore/ #Directory with fastq data

samdir=/data/Tuturuatu/results/sam/ #Sam file output

bamdir=/data/Tuturuatu/results/bam/ #Bam file output

bcf_file=/data/Tuturuatu/results/bcf/ #bcf file output

fq1=_val_1.fq.gz #Read 1 suffix

fq2=_val_2.fq.gz #Read 2 suffix

platform="Illumina"


#trim files before alignment

for file in ${rawdata}*_R1_001.fastq.gz

do

   base=$(basename ${file} _R1_001.fastq.gz)

   /usr/bin/TrimGalore-0.6.6/trim_galore --paired --2colour 20 --basename $base -o ${datadir}
$file ${rawdata}${base}_R2_001.fastq.gz

done

wait

echo "done trimming"


for file in ${datadir}*.txt

do

ls $file >> trimming_reports.txt

done
```

```
#First index the reference genome
time bwa index $ref

#Now, retrieving read group and instrument information.
for samp in ${datadir}*${fq1} #Remember to be explicit with file location
do
        base=$(basename $samp _val_1.fq.gz)
    infoline=$(zcat ${samp} | head -n 1)
    instrument=`echo ${infoline} | cut -d ':' -f1`
    instrumentrun=`echo $infoline | cut -d ':' -f2`
    flowcell=`echo $infoline | cut -d ':' -f3`
    lane=`echo $infoline | cut -d ':' -f4`
    index=`echo $infoline | cut -d ':' -f10`

    Now to incorporate this information into the alignment
    rgid="ID:${instrument}_${instrumentrun}_${flowcell}_${lane}_${index}"
    rgpl="PL:${platform}"
    rgpu="PU:${flowcell}.${lane}"
    rglb="LB:${base}_library1"
    rgsm="SM:${base}"

    echo "Aligning reads for $base" #Be explicit with file location for read 2 and the sam file
output
        time bwa mem -M -R @RG'\t'$rgid'\t'$rgpl'\t'$rgpu'\t'$rglb'\t'$rgsm -t 64 $ref $samp
${datadir}${base}${fq2} > ${samdir}${base}.sam
```

```
    time bwa mem -M -t 64 $ref $samp ${datadir}${base}${fq2} > ${samdir}${base}.sam
    echo "Converting sam file to bam file for $base"
    time samtools view -@ 16 -T $ref -b ${samdir}${base}.sam > ${bamdir}${base}.bam

    echo "Aligning and indexing file"
    samtools sort -@ 16 -o ${bamdir}${base}.aligned.sorted.bam ${bamdir}${base}.bam
    samtools index -@ 16 -b ${bamdir}${base}.aligned.sorted.bam
    rm ${samdir}${base}.sam
done

#chunk bam files for mpileup
ls ${bamdir}*aligned.sorted.bam > ${bamdir}Tuturuatu_bam_list.txt
perl /data/SubSampler_SNPcaller/split_bamfiles_tasks.pl -b ${bamdir}Tuturuatu_bam_list.txt -g
$ref -n 12 -o /data/Tuturuatu/results/chunks | parallel -j 12 {}

#run mpileup on chunks of bam files
for (( i=1; i<=12; i++ )); do
        bcftools mpileup -O b -f $ref -a AD,ADF,DP,ADR,SP -o ${bcf_file}Tuturuatu_${i}_raw.bcf
/data/Tuturuatu/results/chunks/${i}/* &
done
wait
echo “mpileup is done running”

#variant calling on bcf files
for file in ${bcf_file}*.bcf
do
```

```
base=$(basename $file .bcf)

bcftools call $file -mv -O v -o ${bcf_file}${base}_VariantCalls.vcf &

done

wait

echo “variant calling is complete”


#prepare files for filtering with bgzip and indexing

for file in ${bcf_file}*.vcf

do

base=$(basename $file .vcf)

bcftools reheader -s ${bamdir}Tuturuatu_bam_list.txt ${file} -o ${bcf_file}${base}_reheader.vcf

bgzip ${bcf_file}${base}_reheader.vcf

bcftools index ${bcf_file}${base}_reheader.vcf.gz

ls ${bcf_file}${base}_reheader.vcf.gz >> ${bcf_file}list_of_vcf.txt

done


#concatenate the chunked vcf files

bcftools concat --file-list ${bcf_file}list_of_vcf.txt -O v -o

${bcf_file}Tuturuatu_VariantCalls_concat.vcf --threads 16

echo “vcf file is ready for filtering!”
```

<u>Haplotype Construction</u>

```
 #filter bcf file for depth and strand bias at individual sites

setGT Tuturuatu_VariantCalls_filtered.vcf.gz -t q -n . -i ‘FORMAT/DP<5’ ‘FORMAT/SP>60’ >

Tuturuatu_VariantCalls_final_variants.vcf.gz
```

```
#create consensus sequences
#for loop to create consensus files for each individual bam file within the final vcf
for file in /data/Tuturuatu/results/bam/*.aligned.sorted.bam #loop through the bam files
do
#lists the file name to allow for distinguishing individuals
ls $file >> /data/Tuturuatu/results/bcf/Tuturuatu_TLR_consensus.fasta
#creates consensus for haplotype 1
samtools faidx /data/Tuturuatu/ref_genomes/Maui_masurca.fa jcf7180002687310:7857-10280 |
/usr/bin/bcftools-1.11/bcftools consensus
/data/Tuturuatu/results/bcf/Tuturuatu_VariantCalls_final_variants.vcf.gz -s $file -M N -H 1 >>
/data/Tuturuatu/results/bcf/Tuturuatu_TLR_consensus.fasta

#lists the file name to allow for distinguishing individuals
ls $file >> /data/Tuturuatu/results/bcf/Tuturuatu_TLR_consensus.fasta #lists the file name
#creates consensus for haplotype 2
samtools faidx /data/Tuturuatu/ref_genomes/Maui_masurca.fa jcf7180002687310:7857-10280 |
/usr/bin/bcftools-1.11/bcftools consensus
/data/Tuturuatu/results/bcf/Tuturuatu_VariantCalls_final_variants.vcf.gz -s $file -M N -H 2 >>
/data/Tuturuatu/results/bcf/Tuturuatu_TLR_consensus.fasta
done
```

Appendix C: Scripts associated with troubleshooting TLR characterisation in kakī

Script for Extracting Unmapped Reads

```
input=/data/Kaki/results/bam/
work=/data/Kaki/results/unmapped_reads/
spades=/usr/bin/SPAdes-3.15.0-Linux/bin/spades.py


for bam in ${input}*.bam
	do
	base=$(basename ${bam} .sorted.bam)
	echo "Beginning assembly of unmapped paired reads for ${base}..."


	echo "Creating bam files for all paired, unmapped reads in ${base}..."
	samtools view -@ 16 -u -f 13 ${bam} \
	> ${work}${base}_${filename}_unmap_unmap.bam


	samtools sort -@ 32 -n ${work}${base}_${filename}_unmap_unmap.bam \
	> ${work}${base}_${filename}_unmapped.sorted.bam


	echo "Converting bam file to fastq file for ${base}..."
	bamToFastq -i ${work}${base}_${filename}_unmapped.sorted.bam \
	-fq ${work}${base}_${filename}_unmapped.r1.fastq \
	-fq2 ${work}${base}_${filename}_unmapped.r2.fastq
```

```
	echo "Now conducting local assembly for ${base} in region..."
	${spades} \
	--careful \
	-o ${work}assemblies/${base}_${filename} \
	 -1 ${work}${base}_${filename}_unmapped.r1.fastq \
	-2 ${work}${base}_${filename}_unmapped.r2.fastq \
	-m 64
done
```

Script for Aligning Kakī Reads with NextGenMap

```
ref=/data/Kaki/ref_genome/Kaki1_v2.3.fasta #Reference genome for alignment
datadir=/data/Kaki/data/Trimmomatic_Paired_Trimmed/test_alignment/ #Directory with fastq
data
samdir=/data/Kaki/test_align/sam/ #Sam file output
bamdir=/data/Kaki/test_align/bam/ #Bam file output

fq1=_R1.trim.fastq.gz #Read 1 suffix
fq2=_R2.trim.fastq.gz #Read 2 suffix

for samp in ${datadir}*${fq1} #Remember to be explicit with file location
do
	base=$(basename $samp _R1.trim.fastq.gz)
	sudo /data/NextGenMap-0.5.0/bin/ngm-0.5.0/./ngm -r $ref -p --very-sensitive -1 $samp -
2 ${datadir}$base$fq2 -o $samdir$base.sam -t 12
	samtools view -@ 12 -T $ref -b ${samdir}${base}.sam > ${bamdir}${base}.bam
```

```
        samtools sort -@ 12 -o ${bamdir}${base}.aligned.sorted.bam ${bamdir}${base}.bam
        samtools index -@ 12 ${bamdir}${base}.aligned.sorted.bam


done
```

Script for Quantifying Read Depth

```
#Creating 1 kb bed file for script
Kaki.fast.fai awk {“$1”  1 “$2” } > kaki_chromosomes.bed
bedtools makewindows kaki_chromosomes.bed -w 1000 > kaki_chromosomes_1kb.bed


# Counting the number of mapped reads in 1kb windows for along the  whole kaki genome
/data/Kaki/results/bam/*.aligned.sorted.bam
  do
  base=$(basename ${bam} .aligned.sorted.bam)
  while read -r line
    do
    echo "Calculating number of all primary paired reads mapped for ${base} at ${region}..."
    region=$(echo ${line} | awk '{print $1":"$2"-"$3}')
    chr=$(echo ${line} | awk '{print $1}')
    begin=$(echo ${line} | awk '{print $2}')
    end=$(echo ${line} | awk '{print $3}')
    depth=$(samtools view -@ 64 -c -f 3 -F 256 ${bam} ${region})
    echo "${base},${chr},${begin},${end},${depth}" >> /data/Kaki/results/Genome_depth.csv
  done < kaki_chromosomes_1kb.bed
done
#Counting the number of mapped reads within the TLR gene regions
```

```
/data/Kaki/results/bam/*.aligned.sorted.bam
  do
  base=$(basename ${bam} .aligned.sorted.bam)
  while read -r line
    do
    echo "Calculating number of all primary paired reads mapped for ${base} at ${region}..."
    region=$(echo ${line} | awk '{print $1":"$2"-"$3}')
    chr=$(echo ${line} | awk '{print $1}')
    begin=$(echo ${line} | awk '{print $2}')
    end=$(echo ${line} | awk '{print $3}')
    depth=$(samtools view -@ 64 -c -f 3 -F 256 ${bam} ${region})
    echo "${base},${chr},${begin},${end},${depth}" >> /data/Kaki/results/TLR_depth.csv
  done < tlr_regions.bed
done
```

Appendix D: Filtering trials for kakī

Table D.1: Filtering trials for kakī vcf file. The filtering step conducted is shown and resulting SNPs in TLR genes are reported. Abbreviations: maf=minor allele frequency, minQ=minimum quality, minDP=minimum depth, maxDP=maximum depth, TLR=toll-like receptor, SNP=single nucleotide polymorphism.

Filter	TLR 1A SNPs	TLR 3 SNPs	TLR 4 SNPs	TLR 5 SNPs
maxDP < 200	11	0	0	6
minQ > 20	11	0	0	5
minDP > 5	11	0	0	6
maf < 0.5	9	0	0	5
max missing < 0.1	0	0	0	0

Table D.2: Missingness filtering trials for kakī vcf file. Maximum missingness used for filtering is shown (eg. 0.90 = 10% missing data), and resulting SNPs in TLR genes are reported. Abbreviations: maf=minor allele frequency, minQ=minimum quality, minDP=minimum depth, maxDP=maximum depth, TLR=toll-like receptor, SNP=single nucleotide polymorphism.

Max missingness (% missing data)	TLR 1A SNPs	TLR 3 SNPs	TLR 4 SNPs	TLR 5 SNPs
0.9 (10%)	0	0	0	0
0.8 (20%)	5	0	0	5
0.7 (30%)	9	0	0	5
0.6 (40%)	9	0	0	5
0.5 (50%)	9	0	0	5
0.4 (60%)	9	0	0	5
0.3 (70%)	9	0	0	5
0.2 (80%)	11	0	0	6
0.1 (90%)	11	0	0	6