This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Abstract

The advent of next generation sequencing technologies has had a major impact on inference methods for population genetics. For example, community ecology studies can now assess species interactions using population history parameters estimated from genomic scale data. Figs and their pollinating fig wasps are obligate mutualists thought to have coevolved for some 75 million years. This relationship, along with additional interactions with many species of non-pollinating fig wasps (NPFW), makes this system an excellent model for studying multi-trophic community interactions. A common way of investigating the population histories of a community’s component species is to use genetic markers to estimate demographic parameters such as divergence times and effective population sizes. The extent to which histories are congruent gives insights into the way in which the community has assembled. Because of coalescent variance, using thousands of loci from the genomes of a small number of individuals gives more statistical power and more realistic estimates of population parameters than previous methods using just a handful of loci from many individuals.

In this thesis, I use genomic data from eleven fig wasp species, which are associated with three fig species located along the east coast of Australia, to characterise community assembly in this system. The first results chapter describes the laboratory and bioinformatic protocols required to generate genomic data from individual wasps, and assesses the level of genetic variation present across populations using simple summaries. The second results chapter presents a detailed demographic analysis of the pollinating fig wasp, Pleistodontes nigriventris. The inferences were made using a likelihood modelling framework and the pairwise sequentially Markovian coalescent (PSMC) method. The final results chapter characterises community assembly by assessing congruence between the population histories inferred for eight fig wasp species. The population histories were inferred using a new composite likelihood modelling framework. I conclude by discussing the implications of the results presented along with potential future directions for the research carried out in this thesis.
Lay summary

All organisms interact with others in some shape or form. Studying the relationships between sets of interacting species provides an insight into the structure of natural communities. To be able to understand why communities have come to occupy their current distributions, we must consider their history with respect to both their interactions with the environment and other species. Large scale changes at the population level, for example, changes in population sizes due to migrations or local extinctions, leave traces in the genetic material of surviving individuals. Therefore, sampling present day individuals from organisms of interest can enable the population history, i.e. events that affected entire populations, of such species to be reconstructed. For example, such studies have been used to demonstrate how species responded to the severe climatic cycles that have shaped the earth over the last 2 million years. They show that many species in the northern hemisphere were forced to retreat to more southerly locations during the harsh glacial periods but were able to expand again during warmer, more favourable conditions. Knowledge of how species have responded to such climate fluctuations and, in particular, whether sets of interacting species have maintained those relationships during these times, will contribute to our understanding of the processes that have created the high level of current organismal diversity and could aid in predicting how organisms may respond to future climate and habitat changes.

Studying natural communities is challenging because the number of species interacting within them could run to the hundreds or thousands. Herbivorous insects that inhabit plant galls, structures made from various plant tissues within which the insects feed and breed, require these structures to complete their lifecycles and interact with a range of species within them. This makes such species excellent candidates for the study of species interactions within natural communities. In this thesis, I use sets of interacting wasp species, pollinators and their natural enemies, which reside in galls on fig trees along the east coast of Australia to assess the population histories present across a geographically widespread community. The results suggest that, even though the majority of the Australian continent was not impacted directly by glaciers, these species have been greatly affected by past climate fluctuations. However, the sets of interacting species do not show concordant histories through time suggesting a series of complex
responses that were not shaped by a single climatic event but possibly by multiple events that affected the interactions within these communities in different ways.
Declaration

I declare that I am the sole author of this thesis. I conducted all of the analyses and wrote the thesis under the guidance of my supervisors. All work presented in this thesis is my own, with the following acknowledgements:

I was assisted in sample collection by James Cook and Lynsey Bunnefeld. Those samples not collected on this trip were collected by James Cook and Tim Sutton.

I made use of some programming code kindly provided by others:

Jack Hearn provided a number of shell scripts used in the bioinformatic processing pipeline (Chapter 2).

Dominik Laetsch wrote the perl and python scripts used to make the blobplots (Chapter 2).

Lynsey Bunnefeld and Konrad Lohse provided python scripts for block cutting and msprime simulations (Chapters 3 and 4), which I amended accordingly.

Where code was borrowed, the writer has been appropriately acknowledged.

This work has not been submitted for any other degree or professional qualification.

Lisa Cooper
Acknowledgements

I would like to thank my supervisors, Graham Stone and Konrad Lohse, for giving me the opportunity to carry out this project and Ally Phillimore for all the helpful chats and encouragement. Also, I want to say a big thank you to James Cook for introducing me to the figs of Australia, supplying samples and for all your fig related knowledge. In particular, I would like to thank Lynsey Bunnefeld for always putting up with me. Your help, support and encouragement has meant the world and I wouldn’t have made it this far without you. Thank you to Jack Hearn for your help with the bioinformatics and for answering all of my questions, to Dom Laetsch for your help making my blobplots and to James Nicholls for your company in the office and lab expertise over the years.

A massive thank you to Stone lab group members past and present for welcoming me and for all the fun times. I would also like to thank the Cook lab group members for making me feel welcome when I visited and for helping with sample IDs and for kindly answering my questions. To the amazing friends I’ve made in Ashworth throughout this intense process, you guys know who you are and I could not have got to this point without your support and friendship.

Finally, the biggest of thank yous to my family, friends and husband for putting up with me over the last few years. You’ve been so understanding and I hope to be able to make it up to you all for the missed meet ups and gatherings.

I want to dedicate this thesis to Lily and Hedley. You are and will always be my inspiration. You are the most amazing people I will ever have the honour of knowing and I miss you every day. Thank you for always being so proud of me.
# List of abbreviations and model parameters

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>Approximate Bayesian computation</td>
</tr>
<tr>
<td>ADM</td>
<td>Isolation with instantaneous admixture model</td>
</tr>
<tr>
<td>Ahet</td>
<td>Heterozygous sites present in population A</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike information criterion</td>
</tr>
<tr>
<td>ARG</td>
<td>Ancestral recombination graph</td>
</tr>
<tr>
<td>B</td>
<td>Population size scaling parameter ((1/B)*Na)</td>
</tr>
<tr>
<td>BG</td>
<td>Burdekin Gap</td>
</tr>
<tr>
<td>Bhet</td>
<td>Heterozygous sites present in population B</td>
</tr>
<tr>
<td>bp</td>
<td>Base pairs</td>
</tr>
<tr>
<td>bSFS</td>
<td>Blockwise site frequency spectrum</td>
</tr>
<tr>
<td>BVB</td>
<td>Brisbane Valley Barrier</td>
</tr>
<tr>
<td>C.I</td>
<td>Confidence intervals</td>
</tr>
<tr>
<td>cytb</td>
<td>cytochrome b</td>
</tr>
<tr>
<td>diffInL (ΔlnL)</td>
<td>Difference in log likelihoods</td>
</tr>
<tr>
<td>EUKO</td>
<td>Eukobelea samples</td>
</tr>
<tr>
<td>f</td>
<td>Admixture fraction</td>
</tr>
<tr>
<td>Fixed</td>
<td>Fixed differences present between populations A and B</td>
</tr>
<tr>
<td>FON</td>
<td>North <em>Ficus obliqua</em> samples</td>
</tr>
<tr>
<td>FOS</td>
<td>South <em>Ficus obliqua</em> samples</td>
</tr>
<tr>
<td>FRN</td>
<td>North <em>Ficus rubiginosa</em> samples</td>
</tr>
<tr>
<td>FRS</td>
<td>South <em>Ficus rubiginosa</em> samples</td>
</tr>
<tr>
<td>FRSYD</td>
<td>Sydney <em>Ficus rubiginosa</em> samples</td>
</tr>
<tr>
<td>Fst</td>
<td>Difference between total heterozygosity and within population heterozygosity</td>
</tr>
<tr>
<td>FWN</td>
<td>North <em>Ficus watkinsiana</em> samples</td>
</tr>
<tr>
<td>FWS</td>
<td>South <em>Ficus watkinsiana</em> samples</td>
</tr>
<tr>
<td>Gb</td>
<td>Gigabases</td>
</tr>
<tr>
<td>GDR</td>
<td>Great Dividing Range</td>
</tr>
<tr>
<td>hABC</td>
<td>Hierarchical approximate Bayesian computation</td>
</tr>
<tr>
<td>HERO</td>
<td>Herodotia samples</td>
</tr>
<tr>
<td>HV</td>
<td>Hunter Valley</td>
</tr>
<tr>
<td>IM</td>
<td>Isolation with continuous migration model</td>
</tr>
<tr>
<td>kya</td>
<td>Thousand years ago</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>lnL</td>
<td>Log likelihood</td>
</tr>
<tr>
<td>LRT</td>
<td>Likelihood ratio test</td>
</tr>
<tr>
<td>M</td>
<td>Scaled migration rate ((4Nem))</td>
</tr>
<tr>
<td>MIS</td>
<td>Marine Isotope Stage</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood estimate</td>
</tr>
<tr>
<td>MRCA</td>
<td>Most recent common ancestor</td>
</tr>
</tbody>
</table>
\begin{itemize}
\item \textit{mu (\(\mu\))}: Per generation mutation rate
\item \textit{mya}: Million years ago
\item \textit{N50}: Measure of assembly quality (50\% of bases in the assembly are contained in scaffolds greater than or equal to this value)
\item \textit{Na}: Ancestral population size
\item \textit{Ne}: Effective population size
\item \textit{NPFW}: Non-pollinating fig wasp
\item \textit{NSW}: New South Wales
\item \textit{PHIL}: \textit{Philotrypesis} species
\item \textit{pi (\(\pi\))}: Pairwise diversity estimate
\item \textit{POL}: \textit{Pleistodontes} species
\item \textit{PSMC}: Pairwise sequentially Markovian coalescent
\item \textit{QLD}: Queensland
\item \textit{r}: Per generation recombination rate
\item \textit{rho (\(\rho\))}: Population recombination rate ($4Ne$)
\item \textit{SD}: Standard deviation
\item \textit{SFS}: Site frequency spectrum
\item \textit{Shared}: Shared heterozygous sites between populations A and B
\item \textit{SMC}: Sequential Markov coalescent
\item \textit{SNP}: Single nucleotide polymorphism
\item \textit{StLG}: St. Lawrence Gap
\item \textit{SYC}: \textit{Sycoscapeter} species
\item \textit{T}: Time of divergence event (scaled by $2Ne$) – IM parameter
\item \textit{T1}: Time of admixture event (scaled by $2Ne$) – ADM parameter
\item \textit{T2}: Time of divergence event (scaled by $2Ne$) – ADM parameter
\item \textit{theta (\(\theta\))}: Per block diversity estimate
\item \textit{TMRCA}: Time to the most recent common ancestor
\item \textit{WAT}: \textit{Watshamiella} species
\item \textit{ya}: Years ago
\end{itemize}
# Contents

Abstract ........................................................................................................................................ i  
Lay summary .................................................................................................................................. ii  
Declaration .................................................................................................................................... iv  
Acknowledgements .......................................................................................................................... v  
List of abbreviations and model parameters .................................................................................... vi  

1 Chapter 1: Introduction .................................................................................................................. 1  
1.1 Phylogeography ......................................................................................................................... 1  
1.2 Alternative uses for genomic data – detecting selection ............................................................ 4  
1.3 The east coast of Australia: an ideal location for phylogeographic studies ......................... 6  
1.3.1 Barriers to dispersal along the east coast .............................................................................. 7  
1.4 Comparative phylogeography .................................................................................................... 11  
1.5 Studying a community ................................................................................................................ 12  
1.6 Ecologically closed communities of herbivorous insects ......................................................... 16  
1.6.1 An example of an ecologically closed community: figs and fig wasps.............................. 16  
1.6.1.1 Fig development and the lifecycle of a fig wasp ................................................................. 20  
1.6.1.2 *Ficus rubiginosa, Ficus obliqua, Ficus watkinsiana* and their fig wasp inhabitants .......... 22  
1.7 Thesis Outline ............................................................................................................................ 27  
1.7.1 Chapter 2: Whole genome sequencing on a community scale ........................................... 27  
1.7.2 Chapter 3: Profiling the demographic history of *Pleistodontes nigriventris*, the pollinator of *Ficus watkinsiana* .......................................................................................... 28  
1.7.3 Chapter 4: Comparative phylogeography of the inhabitants from a geographically widespread fig wasp community ................................................................. 28  
1.7.4 Chapter 5: General discussion ................................................................................................ 28  

2 Chapter 2: Whole genome sequencing on a community scale .................................................. 29  
2.1 Abstract ....................................................................................................................................... 29  
2.2 Introduction ................................................................................................................................ 30  
2.2.1 Phylogeography using multiple loci ....................................................................................... 30  
2.2.2 Study system and design ......................................................................................................... 33  
2.2.3 Haplodiploidy ....................................................................................................................... 41  
2.2.4 Sample selection .................................................................................................................... 42  
2.2.5 Objectives ............................................................................................................................. 43  
2.3 Methods ...................................................................................................................................... 45  
2.3.1 Sample collection .................................................................................................................. 45
2.3.2 Sample DNA barcoding ................................................................. 45
2.3.3 High-throughput library preparation and sequencing......................... 51
2.3.4 Bioinformatic pipeline overview .................................................. 55
  2.3.4.1 Quality control and processing of sequence reads ......................... 57
  2.3.4.2 Blobplots ........................................................................... 58
  2.3.4.3 Contaminant read removal .................................................... 60
  2.3.4.4 Assembly, mapping, variant calling and repeat masking ............. 60
  2.3.4.5 VCF filtering ....................................................................... 64
2.3.5 Genome-wide concatenated phylogenies ........................................... 64
2.4 Results ....................................................................................... 66
  2.4.1 Selection of individuals for HiSeq sequencing ............................... 66
  2.4.2 DNA quantification and library preparation .................................... 75
  2.4.3 Assessing contamination and assembly/mapping quality ............... 76
  2.4.4 Phylogenetic analysis of concatenated genome-wide SNPs across
      individuals within species ................................................................ 92
2.5 Discussion .................................................................................. 98
  2.5.1 The utility of DNA barcoding ...................................................... 98
  2.5.2 Detecting and removing contaminant reads .................................... 100
  2.5.3 Why does coverage differ between individuals and species? ........... 101
  2.5.4 The meta-assembly approach ..................................................... 103
  2.5.5 Does meta-assembly quality correlate with genetic diversity? ........ 103
  2.5.6 Conclusions .......................................................................... 105
3 Chapter 3: Profiling the demographic history of Pleistodontes nigriventris, the
            pollinator of Ficus watkinsiana .................................................... 107
  3.1 Abstract .................................................................................... 107
  3.2 Introduction ............................................................................... 108
    3.2.1 Habitat barriers and dispersal .................................................. 108
    3.2.2 Australian fig wasp phylogeography ......................................... 109
      3.2.2.1 Ficus watkinsiana and its pollinator, Pleistodontes nigriventris .... 111
    3.2.3 The Burdekin and St. Lawrence Gaps ....................................... 112
    3.2.4 Genomic scale demographic inference ..................................... 113
    3.2.5 Exploring ancestral population size changes using PSMC ............ 114
    3.2.6 Demographic inference using the blockwise SFS likelihood method . 115
    3.2.7 Objectives .......................................................................... 118
  3.3 Methods ................................................................................... 120
    3.3.1 Model fitting .......................................................................... 120
      3.3.1.1 Block cutting ................................................................. 120
3.3.1.2 Data summaries, k-max values and configuration counts ..............121
3.3.1.3 Likelihood analyses of divergence models .................................122
3.3.1.4 Scaling parameter estimates ..................................................125
3.3.1.5 Recombination rate estimation ..............................................126
3.3.1.6 Bootstrapping – correction for LD ..........................................127
3.3.1.7 Model selection – IM vs ADM .................................................128
3.3.2 PSMC analysis ...........................................................................128
3.3.2.1 Variant calling and consensus sequence generation ....................128
3.3.2.2 Running PSMC analyses .........................................................129
3.4 Results .........................................................................................130
3.4.1 bSFS likelihood analyses .............................................................130
3.4.1.1 Isolation with continuous migration ..........................................130
3.4.1.2 Isolation with instantaneous admixture ....................................135
3.4.1.3 Assessing model fit ..................................................................137
3.4.1.4 Model selection – IM vs ADM ...............................................139
3.4.2 PSMC analyses ...........................................................................140
3.5 Discussion .....................................................................................144
3.5.1 Population history of P. nigriventris ..............................................144
3.5.2 Have the Burdekin and St. Lawrence Gaps contributed to the
divergence history of P. nigriventris? ..................................................145
3.5.3 Limitations of the demographic inference methods .......................148
3.5.3.1 Violations of the bSFS scheme assumptions ................................148
3.5.3.2 Are PSMC analyses robust to parameters used? .......................151
3.5.3.3 Generation time ....................................................................152
3.5.4 Conclusions ...............................................................................152
4 Chapter 4: Comparative phylogeography of the inhabitants from a
geospatially widespread fig wasp community ....................................153
4.1 Abstract .........................................................................................153
4.2 Introduction ...................................................................................155
4.2.1 Comparative phylogeography .....................................................155
4.2.2 Model systems for community studies .........................................159
4.2.3 The trophic levels of a fig wasp community ................................159
4.2.4 Ficus rubiginosa, Ficus obliqua and their inhabitants ..................161
4.2.5 East coast Ficus obliqua .............................................................164
4.2.6 Objectives ..................................................................................165
4.3 Methods .........................................................................................168
4.3.1 Pre-processing ............................................................................168
4.3.1.1 Block cutting ................................................................. 168
4.3.1.2 Data summaries, k-max values and configuration counts........ 169
4.3.2 Model fitting .................................................................. 170
4.3.2.1 Composite likelihood analyses of divergence models .......... 170
4.3.2.2 Scaling parameter estimates ........................................... 171
4.3.2.3 Bootstrapping – correction for LD .................................. 171
4.3.2.4 Model selection ........................................................... 172
4.4 Results .............................................................................. 173
4.4.1 Full bSFS vs pairwise composite likelihood – P. nigriventris .... 173
4.4.2 Population pair data summaries ........................................ 176
4.4.3 Composite likelihood analyses .......................................... 180
4.4.3.1 Sycoscapter short, the exception to the rule..................... 184
4.4.3.2 The Sydney population ................................................ 186
4.4.4 The F. rubiginosa community ........................................... 186
4.4.4.1 Pattern of North-South divergence across trophic levels...... 186
4.4.4.2 Presence and direction of gene flow ................................. 190
4.4.4.3 Ne concordance in the North and South across the community ... 190
4.4.5 Philotrypesis (ex F. rubiginosa) vs Philotrypesis (ex F. obliqua) .... 191
4.5 Discussion ........................................................................ 193
4.5.1 Six out of seven members of the F. rubiginosa community diverged in the Pleistocene .......................................................... 193
4.5.1.1 Pleistodentes imperialis and Herodotia ............................... 193
4.5.1.2 The small NPFW .......................................................... 195
4.5.1.3 The wasp inhabitants of F. rubiginosa: a community perspective .. 199
4.5.2 Within-guild comparison of population histories .................... 201
4.5.3 Limitations of the composite likelihood inference method ....... 203
4.5.3.1 Full bSFS vs pairwise composite likelihood inference ......... 203
4.5.3.2 Violations of the composite likelihood assumptions .......... 203
4.5.3.3 Calibrating absolute divergence times ............................. 204
4.5.4 Conclusions ................................................................... 206
5 General discussion ................................................................ 209
5.1 Thesis overview and key findings ........................................ 209
5.1.1 Methodology .................................................................. 209
5.1.2 A phylogeographic overview .......................................... 210
5.2 Study limitations, and what I’d do differently (hindsight is a wonderful thing) ................................................................. 213
5.2.1 Sampling ...................................................................... 213
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.2 Inclusion of selection into demographic models</td>
<td>214</td>
</tr>
<tr>
<td>5.2.3 Bioinformatics</td>
<td>215</td>
</tr>
<tr>
<td>5.2.4 All models are wrong but some are useful</td>
<td>216</td>
</tr>
<tr>
<td>5.3 Future directions</td>
<td>219</td>
</tr>
<tr>
<td>5.4 Concluding remarks</td>
<td>222</td>
</tr>
<tr>
<td>Bibliography</td>
<td>223</td>
</tr>
<tr>
<td>Appendix A – Bioanalyser traces</td>
<td>259</td>
</tr>
<tr>
<td>Appendix B – Library pooling for sequencing</td>
<td>263</td>
</tr>
<tr>
<td>Appendix C – Blobplots</td>
<td>267</td>
</tr>
<tr>
<td>Appendix D – PSMC parameter choice</td>
<td>315</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

1.1 Phylogeography

Phylogeography is the study of the genealogical relationships between lineages (most often populations within species) in geographic space (Hickerson et al., 2010). Avise et al (1987) coined the term to describe a discipline that would bridge the gap between classical population genetics and phylogenetics. Phylogeographic studies aim to answer questions regarding how and when species came to occupy their current distributions, usually in the context of historical biogeographical scenarios. The historical processes believed to shape species’ distributions are often divided into two classes, vicariance and dispersal (Carlquist, 1966; Rosen, 1978). Vicariance biogeography is concerned with the break-up of a species range through a change in the environment (Armstrong, 2011). In contrast, dispersal biogeography is concerned with the movement of species from one area to another, despite the presence of any potential environmental barriers (Armstrong, 2011). To assess the contribution of either process to the spatial and temporal structure of a species, phylogeographic studies often focus on the impacts of biogeographical barriers (such as mountain ranges) and/or major climatic events, e.g. the glacial cycles of the Pleistocene (Bilgin, 2011; Edwards et al., 2016a).

When Avise et al (1987) coined the term, they envisioned the widespread use of mitochondrial DNA (mtDNA) loci in phylogeographic studies of animals. While early studies used non-sequence data such as allozyme allele frequencies to infer spatial population structure, the advent of affordable DNA sequencing technology made it possible to infer population histories from gene phylogenies (gene trees). The commonly adopted approach consisted of building single locus intraspecific gene trees, using sequences for large numbers of individuals sampled over a wide geographic area. Subsequently, any prominent landscape features within the study area were related to the structure in the gene tree (Arbogast and Kenagy, 2001). Several features of mtDNA helped to promote its use in early phylogeographic studies. Due to its maternal inheritance and haploid state, it is not affected by recombination, and has one quarter of the effective population size ($N_e$) of diploid nuclear markers. Lower $N_e$ means that (all other things being equal) fewer generations are required for lineages to sort between geographic regions, resulting in more genetic signal to reveal population structure (Avise et al., 1987; Hickerson et al., 2010). Mitochondrial genes also commonly have a higher and more variable mutation rate (compared to many nuclear genes), resulting in a higher level of
intraspecific polymorphism making it more suitable for inferring population relationships within species (Avise et al., 1987; Nielsen and Slatkin, 2013). Practically, mtDNA was a popular choice for sequencing studies due to its high abundance in cells and, with the increasing availability of Sanger sequencing, it became the marker of choice (alongside the use of chloroplast and nuclear ITS markers in plants) in a wide range of studies (McGuigan et al., 1998; Taberlet et al., 1998; Sullivan et al., 2000; Van Houdt et al., 2005; Koizumi et al., 2014). The increasing availability of mtDNA sequences in sequencing repositories such as Genbank also led to it being considered the ideal tool for animal species identification i.e. sequence barcoding (Hebert et al., 2003).

Despite its historical and ongoing widespread use, several drawbacks have been highlighted that mean interpretations of species and phylogeographical relationships based solely on mtDNA should be treated with caution (Brito and Edwards, 2009). Because mtDNA does not recombine it represents only a single replicate of the coalescent (Kingman, 1982a; Kingman, 1982b) (Figure 1.1), meaning it does not capture any of the variance associated with this stochastic process (Rosenberg and Nordborg, 2002). Extrapolating the structure of a population from a single mtDNA gene tree can be misleading not least because of the randomness associated with generating a single gene tree under the coalescent (Figure 1.2) (Hickerson et al., 2010), but also because mtDNA can be affected by processes (e.g. introgression of mitochondria between species, and selective sweeps caused by maternally co-inherited endosymbionts such as Wolbachia bacteria) that distort the signal of the species demographic history (Hurst and Jiggins, 2005; Nicholls et al., 2012). As a result of these issues, phylogeography is moving towards using gene genealogies to infer phylogeographic informative parameters, such as divergence times and migration rates, under explicit demographic models (statistical phylogeography) as opposed to making inferences directly from gene trees (Knowles, 2009). However, the ability to accurately infer such parameters from data for a single locus is severely limited (Edwards and Beerli, 2000). This is because the ability to discriminate among alternative models comes from the patterns of coalescent (and hence genetic) variation across many loci. For this reason, phylogeographic studies now commonly use data for multiple loci (often incorporating nuclear and mtDNA) (Hope et al., 2011; Guzik et al., 2012; Arias et al., 2014; Aduse-Poku et al., 2015; Mende et al., 2016; Moritz et al., 2016; Potter et al., 2016). However, the ability to utilise multilocus datasets is restricted by the availability of programs that are capable of dealing with larger numbers of loci.
(Cutter, 2013). For example, the isolation with migration models implemented in the program IMa2 can efficiently deal with no more than 200 loci (Hey, 2010) and are frequently used with much less, e.g. 2-12 loci (Camargo et al., 2012; Silva et al., 2015; Menezes et al., 2016; Miranda et al., 2016).

**Figure 1.1:** The coalescent – a genealogy is formed by each lineage randomly ‘choosing’ its parent going back in time. a) The genealogy for a population of 10 haploid individuals tracing the ancestries of 3 sampled lineages back to the most recent common ancestor (black lines). b) The sub-genealogy of the 3 sampled lineages showing the time to the first coalescent event between two of the lineages ($T(1)$) and the time to the second coalescent event ($T(2)$), the ancestor to the 3 lineages.

**Figure 1.2:** The stochastic nature of the coalescent is demonstrated by these 4 genealogical trees that were generated under the standard coalescent model for a sample size of 10. The variation is due to chance alone. (Rosenberg and Nordborg, 2002).
Overcoming this particular computational limitation is an ongoing process that has been spurred on by the rise of high throughput sequencing (Excoffier and Heckel, 2006; Kuhner, 2009; Schraiber and Akey, 2015), which has driven the development of programs able to infer population demographic parameters using genome-scale multilocus datasets (e.g. Gutenkunst et al., 2009; Excoffier et al., 2013). As these genome-wide datasets have become more commonly available, it has been shown that the power to discriminate among alternative population histories comes from better sampling of coalescent variation across many loci, even if from only a few individuals, compared to the traditional approach (described above) of sampling few loci from many individuals (Felsenstein, 2006). Such genome-scale demographic inference approaches are sometimes employed alongside other analyses (e.g. ecological niche models, phylogenies) to build a more comprehensive view of the phylogeographic structure of a species (Derkarabetian et al., 2016; Edwards et al., 2016b; Prates et al., 2016). While the use of large genome-wide datasets in phylogeography is still in its infancy, the continual advances in sequencing and analytical methodology will undoubtedly continue to deepen our understanding of how species distributions have been shaped by biogeographic processes. I consider these issues in greater depth in Chapters 2 and 3. The overall aim of this thesis (expanded below) is to apply genome-scale phylogeographic analyses to a group of insect species that are co-distributed along the east coast of Australia. All inhabit figs, and form part of a fig-centred foodweb. Through this thesis, I use model-based approaches to infer the population histories of individual species, and compare patterns across species to explore the processes of community assembly.

1.2 Alternative uses for genomic data – detecting selection

Levels of genetic diversity differ between species and along the genomes of individuals within species (Romiguier et al., 2014). Determining the factors that influence this level of diversity is a long-standing question within the field of population genetics. Finding the answers will help us to understand how species respond to their environments (Jones et al., 2012; Roesti et al., 2012), how species originate (Nadeau et al., 2012; Nosil and Feder, 2012), and could play an important role in our ability to maintain populations through conservation (Leffler et al., 2012; Ellegren and Galtier, 2016; Ravinet et al., 2017). Selection, drift, mutation and migration all affect inter- and intraspecific diversity levels but the contributions of each process can be hard to disentangle. As a result, the extent to which each of these determines genetic diversity in any single system is usually poorly understood.
(Amos and Harwood, 1998; Ellegren and Galtier, 2016; Ravinet et al., 2017). As an example, genetic data are commonly used to infer population history using explicit demographic models (as discussed above), but the models routinely ignore any impact of selection, assuming neutrality of the loci used. Similarly, analyses that use genomic data to identify signatures of positive selection almost always do not incorporate any effects of population history (Narum and Hess, 2011; Peng et al., 2011). A common method for the detection of selection between populations or species is to use a genome-wide scan (François et al., 2016) to identify hotspots of genetic divergence, and hence to identify candidate loci involved in divergence. Such methods tend to use a ‘sliding window’ approach to plot a measure of genetic differentiation (most commonly Fst) along the genome and look for areas (outlier loci) that show significantly higher Fst compared to background levels in the same region of the genome (Nielsen et al., 2005; Foll and Gaggiotti, 2008; François et al., 2016). These peaks could represent areas of gene divergence where different selection pressures across populations have caused the rise of a particular allele frequency in one population over another. A classic example of such a study is the identification of genes involved in high altitude adaptation in human Tibetan populations (Simonson et al., 2010; Yi et al., 2010). As mentioned, the major drawback of such studies is that they very rarely account for the demographic histories of the populations under study (Keightley and Eyre-Walker, 2007; Frantz et al., 2015; Huber et al., 2016). Ideally both processes, demographic history and selection, should be considered together, and be modelled explicitly in a single framework. However, methods to jointly infer demography and selection are in their infancy. Any analysis that does take selection and demography into account will require long (though not necessarily complete) genome assemblies that are well annotated, because while fragmented genomes can be an excellent source of information on population history (as in this thesis), they do not provide adequate information for either a sliding window approach to identify selection hotspots, or for the annotation of candidate genes. Identification of selective signatures also requires sampling of large numbers of individuals in order to accurately capture the effects of selection on a particular region of the genome and so identify outlier loci.

While inference of selection and population history thus involve similar genomic summary statistics, such as Fst, the sampling strategies required to generate appropriate estimates of these thus differ. As stated above, the aim of this thesis is to use model-based approaches to examine the population histories of a community of fig wasp species. My sampling and analytical framework mean that I can infer
population history with minimal sample sizes using fragmented genome assemblies, but also rule out both the use of a sliding window approach to generate genomic summary statistics and the incorporation of selection into my analysis. I discuss the potential extension of sampling in the fig wasp system to include more complete genome data and selection in Chapter 5.

1.3 The east coast of Australia: an ideal location for phylogeographic studies

Past glacial cycles are known to have played a part in shaping species distributions the world over (Hewitt, 1996; Hewitt, 2000). When glaciers spread, species retreated to areas where conditions were more favourable, known as refugia, or perished (Coghlan et al., 2015). During the warmer interglacial periods species and populations that survived were able to expand beyond these refugial areas (Hewitt, 2004). The process of contraction and expansion can leave patterns in an organism’s genetic structure, which can be investigated to provide clues about how different species respond to severe climate change (Hewitt, 2004). In the northern hemisphere where the Pleistocene (~11.7 thousand years ago (kya) – 2.58 million years ago (mya) (Head and Gibbard, 2015b)) glacial cycles were the most severe, there are well characterised refugial areas which have been the subject of intensive study (Hewitt, 1996; Dumolin-Lapègue et al., 1997; Santucci et al., 1998; Taberlet et al., 1998; Hewitt, 2000; Petit et al., 2002; Nicholls et al., 2010b; Conord et al., 2012; Hearn et al., 2014).

In the southern hemisphere the ice cover did not spread over large areas of the non-Antarctic continents, however the repeated glacial cycles still had dramatic effects on the climate and flora/fauna composition of these regions (Barrows et al., 2002). Australia has been undergoing an intense process of aridification since the opening of the seaway between it and Antarctica at the end of the Eocene (~33 mya) (Martin, 2006). This landmass splitting resulted in steeper climatic gradients from the equator to the poles, the glaciation of Antarctica and the global onset of cooler more seasonal climates (Crisp and Cook, 2013). The shift to cooler climates was complicated in Australia by the northward drift of the Australian landmass to warmer lower latitudes, resulting in warming in the north but continued cooling in the south (Crisp and Cook, 2013). The landscape cover has changed during this period of drying, from predominantly wet rainforest to a mostly arid/semi-arid environment (Martin, 2006). Present day Australia can be separated into three major biomes: the monsoonal tropics in the north, the arid zone across central and western regions, and the mesic regions along the east coast and south-west corner (Catullo and Scott Keogh, 2014). During the Pleistocene glacial/interglacial periods, the
climate cycled between cool/dry and wet/warm conditions respectively (Coghlan et al., 2015). Annual precipitation decreased throughout the Pleistocene with the last glacial period being particularly dry. Rainfall levels in the current interglacial have not returned to the levels of the previous interglacial (Martin, 2006). Throughout these climate cycles, the east coast has become more and more fragmented, with rainforests being separated by mesic habitats such as dry open woodland (Burke et al., 2013). Pollen records indicate the rainforests expanded during the interglacials whereas dry forests and grasses were able to extend their ranges during glacial periods (Kershaw, 1994; Kershaw, 2005). This pattern is supported by studies that use bioclimatic modelling to reconstruct the locations of suitable habitat areas for their species of interest (Hugall et al., 2002; Simpson et al., 2018). For example, Hugall et al (2002) used distribution data and climatic modelling to indicate likely areas of suitable habitat for a rainforest snail. The modelling covers the last 20 kya and shows that the rainforest habitat was severely restricted during the last glacial maximum (LGM) but that it recovers well during the subsequent interglacial (Hugall et al., 2002). Although the modelling only covers the last 20 kya deep sequence divergence found between populations of the snails, and other vertebrates in the region, support the pollen records (Kershaw, 1994; Kershaw, 2005) in suggesting such rainforest contractions and expansions have been occurring throughout the Pleistocene and likely earlier (Hugall et al., 2002). These processes have led to the east coast being subdivided by several known and potential biogeographic barriers, which previous studies have shown to have a major role in population divergence and ultimately speciation in the Australian flora and fauna (Chapple et al., 2011b). This strip of fragmented habitat comprising habitats ranging from tropical rainforest to temperate dry forests therefore makes the east of Australia an excellent latitudinal transect for studying the phylogeographical relationships of interacting species.

1.3.1 Barriers to dispersal along the east coast
The east coast of Australia is characterised by a number of biogeographical zones that are delineated by a range of climate and/or topographical features (Burke et al., 2013). The Great Dividing Range (GDR), Australia’s largest mountain range, runs roughly parallel to the east coast along its entire length (~3500 km) (Smissen et al., 2013) (Figure 1.3). It provides enough elevation change to generate microclimates and enough precipitation to support the mesic habitat along the east coast (Milner et al., 2012). There are eight recognised biogeographic barriers that intersect with the GDR along the east coast and which separate the different biogeographical zones
Biogeographic barriers are here defined, as in Bryant and Krosch (2016), as “areas of unsuitable habitat that interrupt gene flow between populations isolated on either side”. Here I will introduce the four barriers relevant to this study.

(i) The **Burdekin Gap (BG)** separates the northern wet tropics from the dry forests of central Queensland and is located between the towns of Mackay and Townsville (Burke et al., 2013; Bryant and Krosch, 2016) (Figure 1.3). It is the largest dry corridor on the east coast (Bryant and Krosch, 2016) and has been implicated as a barrier to gene flow in a number of taxa by studies using a mixture of mtDNA-only and multilocus datasets (mtDNA only: James and Moritz, 2000; Pope et al., 2001; Schäuble and Moritz, 2001; Nicholls and Austin, 2005; Brown et al., 2006; Edwards...
and Melville, 2010, multilocus: Dolman and Moritz, 2006; Smissen et al., 2013; Bryant and Fuller, 2014). Palaeovegetation and floral evidence to support the timings and extent of the opening of the BG are limited but phylogeographic studies have estimated divergence times for taxa split across the gap (Bryant and Krosch, 2016). For example, Bryant and Fuller (2014) looked at the genetic structure of lineages of a genus of rodent (Melomys) across the BG and other barriers along the east coast. The study assessed genetic structure across the species’ range using eight microsatellite loci, two mtDNA (16S rRNA and cytochrome b (cytb)) loci and a nuclear locus (acid phosphatase type V (AP5)). The authors used a phylogenetic reconstruction, based on the two mtDNA loci and single nuclear locus and calibrated it using dates extracted from a previous phylogenetic analysis to estimate divergence times across the phylogeny. An estimate of 0.99 mya (95% confidence intervals – 0.47 mya – 1.58 mya) was proposed for the time of divergence between the northern and central/southern lineages with the BG as a possible cause. Only a minority of studies have divergence time estimates associated with them, but these broadly agree with this Pleistocene timing, some covering the Pleistocene/Pliocene boundary (Edwards and Melville, 2010; Smissen et al., 2013). Contrastingly, some studies that have assessed genetic structure in species that span the BG have found no evidence of differentiation across the gap (Firestone et al., 1999; MacQueen et al., 2012; Burke et al., 2013).

(ii) The St. Lawrence Gap (StLG) is a second but smaller dry corridor located around 350 km south of the BG between MacKay and Rockhampton (Bryant and Krosch, 2016) (Figure 1.3). As with the BG, the StLG has been referred to in a number of studies as a reason for the patterns of divergence seen in different taxa and, as with those concerning the BG, this evidence is based on a mixture of mtDNA-only and multilocus analyses (mtDNA only: Nicholls and Austin, 2005; Baker et al., 2008; Edwards and Melville, 2010; Chapple et al., 2011b; MacQueen et al., 2012, multilocus: Rix and Harvey, 2012; Burke et al., 2013). A study of Australian glow-worms (Diptera, Arachnocampa) whose distribution encompasses the entire east coast estimated a time of divergence for species either side of the gap at ~6.2 mya (late Miocene) (Baker et al., 2008). The estimate is based on phylogenetic analyses of two mtDNA loci calibrated using an estimated rate of divergence calculated in a previous study of mitochondrial evolution in Heliconius butterflies (Brower, 1994). This Miocene estimate is older than the few other studies that have inferred divergence times in different taxa, which cover the Pliocene/Pleistocene epochs (Chapple et al., 2011b; Potter et al., 2012; Burke et al., 2013). Several studies
Population genomics of a fig wasp community

whose sampling areas cover the StLG find no evidence of divergence across it (James and Moritz, 2000; Schäuble and Moritz, 2001; Brown et al., 2006; Smissen et al., 2013; Bryant and Fuller, 2014), despite some having found evidence for divergence across the BG (Brown et al., 2006; Smissen et al., 2013; Bryant and Fuller, 2014).

(iii) The Brisbane Valley Barrier (BVB) is a large area of lowland open forests surrounded by mountain ranges that spread west to east from the GDR (Bryant and Krosch, 2016) (Figure 1.3). It adjoins the Main, McPherson and Border ranges, which are often the barriers referred to in phylogeographical studies. The higher elevation of the mountains has enabled refugia of closed wet forests to form compared to the dry open forests of the BVB lowlands (Weber et al., 2014). This suggests that the differing habitats present in the area could have presented a barrier to gene flow in different taxa depending on their preferred habitat. The topographical features of this area have been suggested as the barrier to dispersal in a range of taxa inferred using both mtDNA and nuclear loci datasets (mtDNA only: James and Moritz, 2000; Baker et al., 2008; Edwards and Melville, 2010; Chapple et al., 2011a; Chapple et al., 2011b, multilocus: Lucky, 2011; Smissen et al., 2013; Bryant and Fuller, 2014). Three of these studies estimated divergence across the BVB in different reptile species that prefer dry open forests, the timings of each overlapping in the Pliocene (Chapple et al., 2011a; Chapple et al., 2011b; Smissen et al., 2013). As with the previously mentioned barriers, the BVB is not a universally acknowledged barrier to gene flow with some studies reporting no evidence of differentiation across it (Schäuble and Moritz, 2001; Nicholls and Austin, 2005; Brown et al., 2006; Burke et al., 2013).

(iv) The Hunter Valley Barrier (HV) is the last biogeographic break to occur within the sampling region of this study. The HV is a lowland area of dry open woodland located in mid New South Wales near Newcastle (Figure 1.3) (Bryant and Krosch, 2016). It is bordered by two upland regions, the Blue Mountains to the south and the Barrington Tops to the north (Bryant and Krosch, 2016). The HV has been implicated as a barrier to dispersal in two species of skink (Chapple et al., 2011a; Chapple et al., 2011b) despite one of the species’ preferred habitat being dry open woodland like that found within the HV. The timings of the inferred divergences between populations of these two species are noticeably different, one being placed in the Pliocene whilst the other was placed in the Miocene. The majority of studies to infer divergence across the HV are based solely on mtDNA datasets (Eldridge et al., 2001; Moussalli et al., 2005; Eastwood et al., 2006; Baker et al., 2008; Chapple et al., 2011a; Chapple et al., 2011b), with the exception of Hazlitt et al (2014) who use a
combination of mtDNA and microsatellite markers to assess the genetic structure across the HV in rock wallabies. However a range of mtDNA-only and multilocus studies have failed to infer differentiation across the HV (Firestone et al., 1999; James and Moritz, 2000; Nicholls and Austin, 2005; Brown et al., 2006; Lucky, 2011; Burke et al., 2013; Smissen et al., 2013).

Ignoring possible variation due to methodological and calibration differences, there is discordance in estimated divergence times across each of the barriers, and without other evidence (e.g. palaeovegetation data) it is hard to know how long these barriers have affected distributions along the east coast of Australia. However, the small number of studies of species and/or population level divergences along the east coast suggest that these biogeographical barriers have affected different taxa in different ways and to varying extents. These potential barriers to dispersal should thus be considered when investigating divergence events and gene flow among species and populations that inhabit this diverse region of Australia.

1.4 Comparative phylogeography

Comparative phylogeography aims to assess the phylogeographic histories of co-distributed taxa, i.e. it asks whether sets of co-distributed taxa demonstrate shared responses to environmental cues (Arbogast and Kenagy, 2001). Its development as a discipline has followed that of single species phylogeography (see above) in that in its infancy mtDNA was the preferred DNA sequence marker in animals (Avise et al., 1987; Avise et al., 2016). Gene trees were inferred from mtDNA loci for each taxon in the study set and compared to look for concordant structure across species that could be related to the biogeographical history of the sampled region. However these comparisons can be misleading without the additional information that comes from estimating population genetic parameters (e.g. Ne) under an explicit coalescent model (Knowles, 2009). Any differences seen in gene divergences between species could be attributed to the stochastic nature of the coalescent process and/or confounded by the demographic properties of the species under study (Knowles, 2009). A common approach therefore is to estimate the individual population histories of each species in the study from multilocus datasets using a coalescent modelling approach followed by a post hoc comparison of the parameter estimates (Smith et al., 2011; Smith et al., 2012; Satler and Carstens, 2017). This approach accounts for coalescent variance and differences in demography but requires appropriate confidence intervals around the estimates in order to infer statistically significant differences. The other main approach involves estimating parameters
over the combined multispecies dataset under a unified statistical framework, as embodied in programs such as msBayes and MTML-msBayes that use approximate Bayesian computation (ABC) (Hickerson et al., 2007; Huang et al., 2011). This approach has the benefit of statistically inferring the number of taxa that share, for example divergence times, across the combined dataset whilst accounting for differences in other demographic parameters such as population size. However, the specific implementations in msBayes and MTML-msBayes do not allow the parameters associated with each individual species to be inferred. Further, all ABC approaches rely upon summaries of the input data, such that not all of the data are utilised in the inference procedure (Csillery et al., 2010; Robert et al., 2011; Sunnaker et al., 2013). Each approach involves a trade off and therefore the most appropriate one to use will depend upon the specifics of the study in question (e.g. sample availability is likely to be an important consideration as the requirements for each approach can vary widely) (the approaches used in comparative phylogeography will be discussed further in Chapter 4). Comparative phylogeography is an extremely valuable approach that can be applied to answer a range of questions and is particularly promising in the study of natural communities (Hayward and Stone, 2006; Hickerson et al., 2010; Smith et al., 2011; Satler and Carstens, 2017). Understanding whether responses to historical climate events are shared across community members and whether responses affect the interactions between species is vital to our understanding of the processes that have shaped current global biodiversity.

1.5 Studying a community
Community ecology is a well-established discipline and theories regarding community structure and assembly date back to Darwin. Currently there are three main competing viewpoints regarding community assembly processes: niche-based, neutral and historical (Cavender-Bares et al., 2009). The classical viewpoint is the niche-based argument. Early ecologists recognised that a species’ traits could affect its interactions with other species and the environment in a predictable manner (Weiher et al., 2011). On one hand it was reasoned that similar species would share environmental requirements and therefore could potentially co-exist. On the other hand however it was noted that the similarity between species would drive competition and therefore selection towards differing traits, limiting the ability of species to occupy the same niche (Cavender-Bares et al., 2009). This competition-based argument led to the development of the ‘competitive exclusion principle’, which stated that ecologically similar species could not co-exist (Leibold, 1995). The
adoption of this notion led to the thinking that evolutionary processes were not relevant at ecological timescales and an increasing divide between ecology and evolutionary biology. This principle led to the thinking that there were ‘assembly rules’ governing community assembly and that a species could not join a community without the subsequent loss of another species (Weiher and Keddy, 1999). Despite this, evolutionary processes remained the focus of some ecological studies, and more recently phylogenetic relationships of species within communities are being used to investigate the ecological processes that influence their assembly (Weiher et al., 2011). The development of neutral theory followed on from niche-based ideas. Hubbell (2001) challenged the niche-based theory by stating that communities are ever changing, open groups of species whose structure and composition are governed by four processes: speciation, extinction, dispersal and drift (Hubbell, 2001). This view states that species similarity does not predict competition and specialisation, and that interactions with other species or the environment have no effect on community assembly (Cavender-Bares et al., 2009).

The third viewpoint, that historical processes are important for community assembly, was put forward by Ricklefs (1987). He stated that it is important to consider processes on a larger regional scale to help explain community dynamics at the local scale (Ricklefs, 1987). This view brought evolutionary processes back to the forefront of studies on community assembly. More recently the integration of phylogenetics into studies of community assembly has received increasing attention (Webb et al., 2002; Emerson and Gillespie, 2008; Vamosi et al., 2009). The evolutionary relationship between species of a community provides a framework with which to assess both ecological and evolutionary processes that contribute to its assembly (Emerson and Gillespie, 2008). Technological, methodological and computational advances have all helped in the development of new ways to study community processes and to discern what drives them (Johnson and Stinchcombe, 2007).

An important concept when investigating community level processes is the different trophic levels present in all functioning communities. The trophic level of a species is the position it holds within a food web (also known as a trophic interaction network), which is of vital importance when considering the assemblage of more than one species (Montoya et al., 2006). Prior to 1980 the majority of ecological studies focussed on interactions between two trophic levels, primarily plant-herbivore, predator-prey, and parasite-host relationships, the bi-trophic perspective (Mooney et al., 2012). Such studies concentrated on the bi-trophic niche concept,
Population genomics of a fig wasp community

which hypothesises species richness within a community (e.g. herbivore numbers) to be limited by the availability of feeding sites (e.g. host plants) (Singer and Stireman, 2005). However a paper published in 1980 by Price et al suggested that a more realistic way to approach such studies was to consider a third trophic level, in this case the natural enemies of herbivores (Price et al., 1980). This tri-trophic perspective combines both ‘bottom-up’ and ‘top-down’ effects between trophic levels, and suggests that herbivore richness depends upon the diversity of ways in which the herbivore is able to use its host plant not only for food (bottom-up) but also for defence from enemies (top-down) (Elbanna, 2007). These concepts are important to consider when contemplating sets of species at different trophic levels as they can help form hypotheses of different possible scenarios. For example Singer and Stireman (2005) state that the biggest effect on herbivore diversification when considering the tri-trophic concept over the bi-trophic one is the acquisition of ‘enemy free space’ (meaning spatial or temporal niches in which they can successfully evade enemy attack) by the herbivores.

The movement towards an accepted cross over in the disciplines of community ecology and evolutionary biology has highlighted the importance of considering the interactions between community members at the population level (Hayward and Stone, 2005; Johnson and Stinchcombe, 2007; Emerson and Gillespie, 2008; Weiher et al., 2011). The extent to which population histories are shared across sets of interacting species through space and time will play a role in shaping a community into its current structure, and is crucial for assessing the extent to which co-evolution (reciprocal evolution of character states in two or more species) is to be expected (Cavender-Bares and Wilczek, 2003). Understanding the evolutionary processes involved in the assembly of natural communities is crucial to the development of future responses to human induced changes (e.g. climate change, biological control of invasive pests and conservation based species re-introductions) (Sax et al., 2007; Zvereva and Kozlov, 2010; Zhu et al., 2015). Assessing the process of community assembly in terms of the population histories of members sampled from different trophic levels (e.g. host and parasite) has led to the formation of three hypotheses that describe the expected patterns under different scenarios (Stone et al., 2012; Andrew et al., 2013) (Figure 1.4): 1) co-diversification – host and parasite are expected to disperse through space and time together leaving a pattern of concordant histories (and the potential for co-evolution) (Figure 1.4(a)), 2) host-shift – new parasites enter the community following dispersal of the host leaving the potential for discordant patterns of population histories across space and time.
(Figure 1.4(b)), and 3) host-tracking – the parasite is expected to track the host through space and time leaving a period of enemy free space for the host and population histories that are concordant in space but discordant in time (Figure 1.4(c)) (Hougen-Eitzman and Rausher, 1994; Hoberg and Brooks, 2008; Nicholls et al., 2010a; Phillips et al., 2010; Smith et al., 2011). The testing of these hypotheses is particularly suited to studying the species-rich communities of herbivorous insects and their natural enemies, in which all three hypotheses have been supported at different trophic levels and in a range of systems (Lewis et al., 2002; Stireman and Singer, 2003; Elbanna, 2007; Bailey et al., 2009; Stone et al., 2012).

![Figure 1.4](image)

**Figure 1.4:** Population history diagrams detailing three hypotheses of community assembly involving two species (e.g. host (black) and parasite (red)) from four populations. a) co-diversification – the two species spread through space and time in the same direction (arrows) from a shared origin creating concordant histories. b) host-shift – the two species spread in different directions resulting in discordant histories suggesting they could have been exposed to each other for differing lengths of time. c) host-tracking – the two species spread at different times in the same direction resulting in concordant topologies but with one species lagging behind the other in time.
1.6 Ecologically closed communities of herbivorous insects

Herbivorous insects and their natural enemies, along with their host plants, contribute a large proportion of the planet’s biodiversity (Price, 2002). These systems have become the focus of a huge array of ecological and evolutionary studies. One aspect of such systems that makes them particularly useful for assessing multltrophic interactions is that the majority of insect herbivores are specialised feeders (Stamp, 2001). This applies particularly to herbivores (such as gall inducers and leaf miners) that live within plant tissues. Because the sets of species in such specialist communities do not interact with other communities, i.e. they are ecologically ‘closed’, they can meaningfully be considered in isolation of other species and communities (Askew, 1980; Hayward and Stone, 2005; Bihari et al., 2011; Leppanen et al., 2013), making them popular subjects for ecological study (Lewis et al., 2002; Bailey et al., 2009). The tight association with their host plants, and in particular their associated plant structures (i.e. galls, leaves), also tends to make them easier to sample and can help to obtain distributional data. These communities also have additional trophic levels of specialised natural enemies, particularly insect parasitoids (Hayward and Stone, 2006; Herre et al., 2008; Smith et al., 2011; Segar et al., 2012; Kaliszewska et al., 2015; Nair et al., 2016). Parasitoids are thought to represent a higher proportion of insect species compared to their herbivorous hosts (Price, 2002; Hrcek et al., 2013) and they play important roles in regulating herbivore populations (Bailey et al., 2009). This thesis will focus on one such system - the galling wasps that pollinate fig trees and their associated parasitoids.

1.6.1 An example of an ecologically closed community: figs and fig wasps

The genus *Ficus* (Moraceae) contains more than 750 species, making it one of the most speciose genera of land plants (Cook and Rasplus, 2003; Ronsted et al., 2008a). Figs show a pan-tropical distribution and due to their ability to fruit year round they represent keystone species for numerous frugivores across this range (Shanahan et al., 2001). It has also been noted that they have played significant roles in forest recolonisation (Harrison, 2005; Herre et al., 2008), suggesting they could be important in areas affected by deforestation. Figs show remarkable variation in growth form both within and across species (Dixon, 2003; Harrison, 2005). They are able to grow as free standing trees, shrubs, hemi-epiphytes and lithophytes and are found in a range of habitats from rainforests to rocky cliff faces (Ronsted et al.,)
Population genomics of a fig wasp community

2008b; Darwell, 2013). There are two breeding systems present across the genus *Ficus*, monoecy and dioecy, with approximately 50% of species showing each respectively (Cook and Segar, 2010). Male and female function is carried out within the same figs of monoeicous species whereas they occur on separate trees in dioecious species (Herre et al., 2008). One of the most interesting aspects of fig biology and the main reason they have been so heavily studied is the mutualistic relationship they share with the Agaonidae wasps (Hymenoptera, Chalcidoidea) that pollinate their flowers.

Figs and their pollinating fig wasps have been used extensively in studies investigating co-evolutionary relationships (Wiebes, 1979; Machado et al., 2005; Ronsted et al., 2005; Cruaud et al., 2012; McLeish and Van Noort, 2012; Wachi et al., 2016). They are a classic example of an obligate mutualism in which each partner relies entirely on the other to ensure their respective reproductive success (West and Herre, 1994; Herre et al., 2008). Although originally thought to be a specific ‘one-to-one’ relationship (i.e. each fig tree has a single specific pollinator), it is now known that this is not always the case (Haine et al., 2006; Chen et al., 2012; McLeish and Van Noort, 2012; Darwell et al., 2014; Yang et al., 2015; Wachi et al., 2016; Sutton et al., 2017), however the relationship between host(s) and pollinator(s) is still highly conserved (Cruaud et al., 2012). This is demonstrated by a large phylogenetic study that sampled nearly 200 interacting pairs of fig and pollinator wasp species using several nuclear and organelle genes (Figure 1.5) (Cruaud et al., 2012). The study suggests the mutualism has been co-evolving for ~ 75 million years (Cruaud et al., 2012). Currently, the widely accepted view of diversification within this ancient relationship is one of co-divergence at higher taxonomic levels (i.e. fig sections and wasp genera) with frequent fig host-switching occurring amongst closely related wasp taxa (Weiblen and Bush, 2002; Machado et al., 2005; Silvieux, 2008; Cook and Segar, 2010; Cruaud et al., 2012; McLeish and Van Noort, 2012). The high degree of specificity of pollinators to their hosts is a property shared amongst many of the other non-pollinating wasp inhabitants of figs (West et al., 1996). It is this element of the system alongside the fact that the fig fruit is central to the completion of the wasps’ lifecycles that make figs and their associated fig wasps excellent model systems for community level studies (Herre et al., 2008; Ghara and Borges, 2010; Segar et al., 2013; Segar et al., 2014). Despite being the focus of so much attention the majority of fig/fig wasp related studies have concentrated on figs in Africa, Asia and the Americas (Compton, 1993; Compton, 1994; West and Herre, 1994; West et al., 1996; Compton et al., 2009; Kobmoo et al., 2010; Chen et al., 2012; Cornille et al.,
Population genomics of a fig wasp community

2012; Vieira et al., 2015), leaving the Australian Ficus species relatively understudied.
Population genomics of a fig wasp community

Figure 1.5: BEAST phylogenies showing the evolutionary history of figs and their pollinating fig wasps. Figure from Cruaud et al (2012). The colours represent groups of figs and their associated genera of pollinators and the pie charts show the likelihood of different geographic areas of origin. The section Malvanchera clade includes the three figs relevant to this study with their associated pollinators being found in the Pleistodontes clade.

Throughout this thesis, I will use the term ‘fig wasp’ to refer to any wasp that develops within the fruit of fig trees. Reference to the different trophic roles of these wasps will be specified accordingly.

1.6.1.1 Fig development and the lifecycle of a fig wasp

As stated above, all fig trees rely on a mutualistic relationship with pollinating fig wasps to pollinate their flowers, located within the enclosed inflorescence (syconium) characteristic of figs (Silvieus, 2008; Segar et al., 2014). In return the fig wasp relies on the fig fruits to provide oviposition sites and a larval food source following egg hatching (Silvieus, 2008). The lifecycle of the pollinating fig wasps is tightly linked to the developmental stages (A-E) of fig growth (Figure 1.6) (Galil and Eisikowitch, 1968; Weiblen, 2002). The focus from here on in will be on fig and pollinator development in monoecious fig species. Small, immature figs (stage A) loosen as they grow, releasing volatile organic compounds that attract pollen-laden female pollinating fig wasps (receptive B stage) (Borges et al., 2008; Silvieus, 2008; Ghara and Borges, 2010). These female wasps enter the fig through a specialised entrance, named an ostiole, often losing their wings and antennae in the process (Cook and Rasplus, 2003; Dunn et al., 2008). Inside the syconium the wasps lay eggs (oviposit) in some female flowers whilst pollinating others (Cook and Lopez-Vaamonde, 2001). The flowers that receive an egg form galls in which the larvae develop, while those that do not, develop into seeds (Herre et al., 2008). This development occurs during the interfloral stage (stage C) (Ghara and Borges, 2010). Once the larvae are fully developed, they pupate, and the male wasps emerge into the fig cavity and mate with females that are still enclosed in their galls (Harrison, 2005; Herre et al., 2008). Following mating the females emerge, collect pollen (either actively or passively dependent upon the species) and then exit the fig via a hole chewed through the fig wall by the males (stage D) (Cook and Lopez-Vaamonde, 2001; Ghara and Borges, 2010). The females, who live from a few hours to a couple of days, must locate another receptive fig to repeat the cycle (Harrison, 2005; Sutton et al., 2016). The males usually die within the syconium (Sutton et al., 2016). Once the wasps have emerged, the figs ripen, changing colour and smell in order to
attract frugivorous seed dispersers (stage E) (Borges et al., 2008; Ghara and Borges, 2010). The requirement to maintain pollinator populations means fig trees fruit all year round, with asynchrony in fruiting at the population level allowing pollinators to find receptive figs upon emergence (Shanahan et al., 2001; Weiblen, 2002). The interfloral stage (stage C) is expected to last approximately 4-6 weeks (Harrison, 2005; Jia et al., 2008), however this timing can be highly variable (3-20 weeks) (van Noort, S. & Rasplus, JY. 2017. Figweb: figs and fig wasps of the world. URL: www.figweb.org.za (accessed on 27/05/17)). The lifecycle of pollinating fig wasps can lead to high levels of sib-mating depending upon the number of females that enter each syconium, which is likely to vary locally and across fig species (Herre et al., 2008; Greeff et al., 2009; Sutton et al., 2016). For example, fig trees with larger syconia can support larger numbers of females, which could result in a reduction of sib-mating compared to fig trees with smaller syconia.

Figure 1.6: The 5 stages of monoecious fig development (A-E) and the associated lifecycle of pollinating fig wasps. Figure from Harrison (2005). Syconia become receptive and attract female pollinators during B-stage. The pollinators enter the fig, pollinate female flowers whilst laying their eggs in others. The flowers that do not receive an egg develop into seeds, those that do develop into galls within which the fig wasp larvae grow (C-stage). After a period of 4-6 weeks, the males emerge, mate with the females and chew an exit hole through the fig wall to allow the pollen-laden females to disperse (D-stage). E-stage figs ripen and are consumed by frugivores that subsequently disperse the seeds. Parasitic non-pollinating wasp species attack the figs during B- and C-stages.
The fig-fig wasp system also supports other wasps, referred to as non-pollinating fig wasps (NPFW), which either gall the fig without pollinating it (wall or ovule gallers) or develop as parasitoids of other fig inhabitants (Silvieus, 2008). Each fig species can support up to 30 NPFW species, adding an interesting third trophic layer to this closed community (Cook and Rasplus, 2003). This group of NPFWs can be classified into four categories, as stated in Segar et al. (2013): 1) Large gallers (i.e. bigger than the pollinators) whose larvae gall flowers or wall tissue; 2) Large parasitoids of the large gallers, which either take over the galls of other wasps or feed on other wasps directly; 3) Small gallers (a similar size to pollinators) whose larvae gall flowers or consume seeds; 4) Small parasitoids of pollinators or other small gallers. These groupings highlight the fact that wall gallers and their associated parasitoids represent a separate trophic compartment within the fig. This contributes an additional set of interacting species to consider when studying these diverse systems along with the added interest of the effects such interactions have on the mutualistic relationship of figs and fig wasps. The reason(s) why this mutualism has been maintained and pollinators do not over exploit the figs is an ongoing area of research (Herre et al., 1999; Cook et al., 2004; Weiblen, 2004; Dunn et al., 2008; Herre et al., 2008). One view suggests a role for parasitoids in ‘policing’ the mutualism (Dunn et al., 2008). Pollinators preferentially gall flowers furthest from the outer fig wall (i.e. closest to the centre of the fig) whereas those closest to the fig wall become seeds (Al-Beidh et al., 2012). Many parasitoids attack figs from the outside through the fig wall allowing pollinators the chance to escape these attacks by only galling these inner flowers. This suggests that by avoiding their enemies in this way pollinators have also benefitted the fig by leaving some flowers free to form seeds (Dunn et al., 2008). Investigation of the biology and species level interactions of these NPFW has been severely neglected compared to those concerned with the pollinating fig wasps (Borges, 2015). However, as they receive more and more attention such studies will help to shed light on the roles NPFW species play in these complex multitrrophic communities.

1.6.1.2 *Ficus rubiginosa, Ficus obliqua, Ficus watkinsiana* and their fig wasp inhabitants

Australian fig-fig wasp interactions are the least studied within the fig-fig wasp literature (Segar et al., 2014), and their biogeography provides an opportunity to explore impacts on population structure, divergence and speciation. The focus of this thesis will be the wasp inhabitants associated with three *Ficus* species found
predominantly on the east coast of Australia (Ficus watkinsiana, Ficus rubiginosa and Ficus obliqua). All three are pollinated by wasps of the genera Pleistodentes and showcase an array of NPFW, most genera of which are also shared across the three fig species. Both F. watkinsiana and F. rubiginosa are endemic to Australia, whereas F. obliqua is also distributed throughout the Pacific Islands (Dixon et al., 2001; Dixon, 2003). Ficus watkinsiana is a rainforest specialist that is found in two distinct populations in eastern Queensland (Figure 1.7) (Lopez-Vaamonde et al., 2002). The figs of F. watkinsiana are large and are often found high up in the tree canopy, which can make sampling challenging (Figure 1.8). It is currently designated as being pollinated by a single species of pollinating fig wasp, P. nigriventris (Lopez-Vaamonde et al., 2002; Dixon, 2003).

Figure 1.7: The distribution of Ficus watkinsiana. Figure from Dixon (2003).
Sister species, *F. rubiginosa* and *F. obliqua*, are habitat generalists that are able to grow as free standing trees, hemi-epiphytes or lithophytes and are found distributed across the majority of the east coast (Figures 1.9 and 1.10) (Dixon et al., 2001; Darwell, 2013). They produce small to mid-sized figs that support a diversity of wasp inhabitants (Segar et al., 2014) and due to their distributions (Figures 1.9 and 1.10) and varied growth forms (Figure 1.8) are easy to sample. *Ficus obliqua* and *F. rubiginosa* have been reported as being pollinated by two and five, pollinator species across their ranges, respectively (Lopez-Vaamonde et al., 2002; Darwell et al., 2014).
Figure 1.9: The distribution of *Ficus rubiginosa*. Figure from Dixon (2001).

Figure 1.10: The distribution of *Ficus obliqua*. Figure from Dixon (2001).
Population genomics of a fig wasp community

The pollinators of *F. obliqua*, *P. xanthocephalus* and *P. greenwoodi*, have not been extensively sampled and so their respective distributions are not very well characterised. When they were first categorised as two species, *P. xanthocephalus* had only ever been found on trees north of Cairns whereas *P. greenwoodi* was found in all other sampled locations (Lopez-Vaamonde et al., 2002). However since then, *P. xanthocephalus* has been sampled as far south as Townsville and *P. greenwoodi* from north of Cairns (J. Cook, personal communication), implying that there is a large overlap in their distributions.

The story of the *F. rubiginosa* pollinators is even more complex, a recent study suggesting this *Ficus* species is actually pollinated by a complex of five sister species (*P. imperialis* sp. 1-5), having originally been classified as a single pollinator species across the hosts range (Darwell et al., 2014). The distribution of these five species is noteworthy in that a single species appears to dominate in the New South Wales portion of the host’s range, whereas in Queensland all five species are present in differing proportions according to location (Figure 1.11) (Darwell et al., 2014).

![Figure 1.11: The distribution and frequencies of the five *Pleistodontes imperialis* species that pollinate *Ficus rubiginosa*. The numbers indicate the sample size of each region. Figure from Darwell (2013).](image-url)
There is a wide range of NPFW associated with these fig hosts. They represent multiple trophic levels covering small and large gallers and several parasitoids. The most abundant species are the small parasitoids of the genera *Sycoscapter*, *Philotrypesis* and *Watshamiella* (subfamily *Sycoryctinae*) (Segar, 2011). Wasps of the genus *Sycoscapter* are known to be parasitoids that directly attack pollinator larvae whereas *Philotrypesis* wasps are thought to be inquilines that lay their eggs in the galls of the pollinators, which can in some cases result in the death of the pollinator larvae (Segar, 2011; Darwell, 2013). *Watshamiella* wasps are hyper-parasitoids, meaning that they use the holes drilled by parasitoids to attack the parasitoid larvae (in this case the *Sycoscapter* parasitoids) (Compton et al., 2009; Segar, 2011; Darwell, 2013). The galling wasps, particularly the large ones, are found at lower densities, meaning that they can be harder to sample (Segar, 2011; Darwell, 2013). Genera in common across these fig hosts include *Herodotia*, *Eukobelea* and *Pseudidarnes*, which are members of the subfamilies *Epichrysomallinae* (*Herodotia*) and *Sycophaginae* (*Eukobelea* and *Pseudidarnes*) (Darwell, 2013). These gallers exploit the fig syconium to provide egg laying sites, like the pollinators, however, they do not provide the fig with any service in return and so represent an additional layer of complexity in these fig based communities.

### 1.7 Thesis Outline

This thesis aims to utilise the advances in high throughput sequencing and methodologies capable of analysing genome-level datasets to contribute to our understanding of the biogeography of the east coast of Australia and the structure of fig wasp communities along this latitudinal transect. The following sections provide brief overviews and the aims of each chapter in this thesis.

#### 1.7.1 Chapter 2: Whole genome sequencing on a community scale

In Chapter 2 I use DNA barcoding to identify and select appropriate individuals of 9 fig wasp species for use in preparation of individual-level genomic libraries and high throughput genome sequencing. The aim of this chapter was to sample members of Australian fig wasp communities as widely as possible, and to generate low coverage genomic datasets for multiple individuals of each species from populations spanning known biogeographic barriers along the east coast. I outline the laboratory and subsequent bioinformatic processes involved in generating individual wasp genome-wide datasets and end by evaluating the success of the sampling and assessing potential biases across the multispecies dataset.
1.7.2 Chapter 3: Profiling the demographic history of *Pleistodontes nigriventris*, the pollinator of *Ficus watkinsiana*

In Chapter 3 I use a genome-wide dataset to infer the demographic history of the pollinating fig wasp, *Pleistodontes nigriventris*. *Pleistodontes nigriventris* pollinates the rainforest fig tree *Ficus watkinsiana*, which is found in two disjunct populations in Queensland, Australia that are separated by two well known biogeographic barriers, the Burdekin and St. Lawrence Gaps. I employ two approaches to infer the population history of this wasp across these divides. The aim of this chapter was to reconstruct the demographic history of *P. nigriventris* using minimal numbers of sampled individuals and low coverage genome-wide data, and to relate the resulting history to the presence of known biogeographic barriers and what is known of Australia’s climate history.

1.7.3 Chapter 4: Comparative phylogeography of the inhabitants from a geographically widespread fig wasp community

In Chapter 4 I infer population histories for eight species of fig wasp from the widespread generalist fig tree species *Ficus rubiginosa* and *Ficus obliqua*. The fig wasp species include multiple trophic levels and the sampling spans four known biogeographic barriers along the east coast of Australia. The aim of this chapter was to infer the demographic history of each species from genome-wide datasets and to interpret the results in the context of the structure of these enclosed multitrophic communities. A composite likelihood inference approach that allows the calculation of confidence intervals around all of the parameter estimates to aid in between-species comparisons was used to infer the demographic history of each species.

1.7.4 Chapter 5: General discussion

This chapter provides an overview of the results of the thesis, discusses the important lessons learned and presents ideas of future directions for this research.
Chapter 2: Whole genome sequencing on a community scale

2.1 Abstract
The technological advances that led to high throughput sequencing have revolutionised the disciplines of phylogeography and population genetics, along with many more. The confirmation that with increased numbers of loci comes greater power to infer the genealogical histories of populations has meant genomic level studies are increasingly prevalent in the phylogeographic/population genetic literature. Numerous challenges still present themselves when considering the use of high throughput sequencing in a project however. One challenge is the ability to obtain high enough quantities of DNA from small organisms for the preparation of sequencing libraries. This issue can be overcome by pooling the DNA from several individuals, however this is not suitable for use in population analyses that use small numbers of individuals per population. Massive improvements in library preparation methods have allowed smaller starting quantities of DNA to be effectively used but obtaining even these amounts can be difficult from some taxa. Another challenge is producing a reference sequence for the taxon of interest if one is not already available, which is the case for the majority of non-model organisms. The wasps associated with fig trees across their pan-tropical distribution form interesting ecologically closed communities with which to address a range of questions relating to shared population responses across space and time. The difficulty of sampling some of these species, either because they are present at low abundances across their distributions or due to the inaccessibility of the trees they reside in, means an approach that can utilise population samples of very small numbers of individuals would be most appropriate in this instance. However the tiny size of these wasps presents a practical hurdle to overcome in order to produce whole genome datasets for such an approach. Here I present the process used to generate individual level low coverage genome-wide sequence data for eleven species of fig wasp from three fig trees in Australia. Each individual level dataset was combined per species and assembled to provide reference sequences for the downstream processing. The sampling design was chosen to enable the use of these datasets in a series of population genomic analyses (Chapters 3 and 4).
2.2 Introduction

2.2.1 Phylogeography using multiple loci

Phylogenetic relationships, and the methods used to infer them, are fundamental tools in evolutionary biology (Pace et al., 2012). When phylogenetic analyses address deep evolutionary splits between lineages, each sampled species (or group) is typically represented by one or a small number of individuals, under the assumption that variation within species is small compared to differences between them (Brito and Edwards, 2009). In contrast, assessing the evolutionary relationships between populations within a species or between closely related species requires a different approach (Brito and Edwards, 2009). Intraspecies phylogeography is concerned with establishing population scale relationships across geographic space (Avise et al., 1987). Studies within the discipline aim to answer such questions as: Is genetic variation within a species uniformly distributed over space, or are there discontinuities that separate discrete populations? Do such discontinuities correspond to specific physical or other environmental features? To what extent do relationships between populations reflect the evolutionary history of species dispersal? The answers to such questions provide fascinating insights into how species have responded to environmental change and the complex process of speciation. The traditional approach to such studies was to use sequence for mitochondrial DNA (mtDNA) loci sampled from large numbers of individuals to infer intraspecies histories (Avise et al., 1987). This approach was based on the wide availability of Sanger sequencing technology (Sanger et al., 1977) and of conserved primer sets amplifying loci such as cytochrome b (cytb) and cytochrome c oxidase (CO1) that often contain moderate to high within-species polymorphism. These tools make it possible to sequence fragments (up to ~1000bp) of a single locus in numerous individuals on a reasonable budget (e.g. 100 individuals sequenced in both directions is <£500).

However, investigating phylogeographical relationships using only mtDNA loci is essentially basing inference on a single gene history. MtDNA does not recombine and therefore all the genes that make up the molecule share the same history (i.e. are a single replicate of the coalescent process) (Kingman, 1982a; Kingman, 1982b). The coalescent is a stochastic model that describes the distribution of genealogical relationships that trace back in time to the most recent common ancestor (MRCA) of the sampled sequences (a genealogy). The use of multiple loci to infer within species relationships allows more of the variation present in this stochastic process to be captured (Hein et al., 2005). Due to recombination, the nuclear genome is a mosaic
of genealogies that, unlike the gene specific effects of selection, will all have experienced the same demographic history (Hudson, 1990). This means that if the loci sampled are neutral and unlinked, each can be thought of as an independent replicate of the coalescent process. Thus the distribution of genealogies across loci contains information about the demographic history of the population (Hein et al., 2005). An extremely useful feature of the coalescent is that it describes the genealogy for \( n \) sampled genes and thus it is not necessary to know the genealogy for the whole population (in contrast to modelling changes in allele frequencies forward in time) (Felsenstein, 2003). This greatly increases computational efficiency when simulating genealogies under the coalescent and has made it possible to derive likelihoods under the coalescent for alternative population histories. The rate of the exponential waiting times between coalescent events is proportional to the number of possible pairs of lineages at each interval, consequently, most of the history linking a sample of gene copies (i.e. the total height of the genealogy, time to most recent common ancestor, TMRCA) is expected to be spent waiting for the final coalescent event to occur (i.e. the last two gene copies before the MRCA is reached) (Hein et al., 2005). This means that on the coalescent timescale the majority of coalescent events occur quickly (i.e. in the recent past) near the tips of the genealogy. It has been shown that increasing the number of gene copies sampled (i.e. individuals sampled for a given gene) causes a disproportionately small increase in the TMRCA and the total branch length of the tree as the extra copies are expected to add only short external branches at the tips of the tree (Hein et al., 2005). In contrast, greater power to discriminate among alternative population histories comes from better sampling of coalescent variation across many loci, even if from only a few individuals, compared to the traditional approach of sampling few loci from many individuals for inferring intraspecies histories (Felsenstein, 2006). An example of the extreme end of minimal sampling is the approach developed by Li and Durbin (2011) with which human demography can be reconstructed from the sequence of a single individual.

Assuming that the \( n \) sampled loci are selectively neutral, the genealogical process can be separated from the mutation process (Hein et al., 2005). Under this assumption the probability that a gene is passed on to the next generation is equal across all genes sampled. To infer the population history for a given sample, a distribution of genealogies is derived under the specified coalescent model and mutations are mapped onto these assuming a mutation model (i.e. a model that specifies the rate at which mutations cause changes to the DNA sequences sampled
Population genomics of a fig wasp community

e.g. the infinite sites model (Nielsen and Slatkin, 2013). The probability of the observed data can be calculated as a function of the parameters given the specified underlying demographic history (Hein et al., 2005). For example, using likelihood theory, the combination of parameter values that maximises the likelihood of the observed data under the specified model corresponds to the most probable history for the sample given the available information. The basic coalescent assumes a sample from a single population reproducing under the Wright-Fisher model. This commonly used model of reproduction assumes a haploid population with discrete, non-overlapping generations in which gene copies are transmitted by random sampling of parents in the previous generation (Felsenstein, 2003; Hein et al., 2005; Nielsen and Slatkin, 2013). However, coalescent theory has been extended to encompass more complex and realistic histories involving population divergence, population structure and changes in effective population size through time (Nielsen and Wakeley, 2001; Hein et al., 2005; Nielsen and Slatkin, 2013).

The ability to produce large multilocus datasets has been expanded enormously by the huge advancements in sequencing technologies since the birth of Sanger sequencing in the 1970s (Sanger et al., 1977). There are now a range of available methods that enable the generation of genome-wide data. These include whole genome sequencing and ‘genomic reduction’ techniques such as restriction-site associated DNA sequencing (RADSeq) (Baird et al., 2008), transcriptome sequencing (e.g. RNAseq (Wang et al., 2009)) and targeted enrichment (Mamanova et al., 2010). ‘Genomic reduction’ techniques aim to reduce sequencing costs by sub-sampling the genome whilst still producing datasets for hundreds or thousands of loci (Nicholls et al., 2015). RADSeq sequences short fragments from sites defined by restriction enzymes and results in large single nucleotide polymorphism (SNP) datasets (Davey and Blaxter, 2010). RAD data have been used to find quantitative trait loci (QTL) (loci that are associated with the variation in a particular trait) (Chutimanitsakun et al., 2011), assess phylogenetic relationships (Wagner et al., 2013) and investigate phylogeographic structure (Emerson et al., 2010). RADSeq is an effective method for use with SNP based inference methods but is currently limited in its applicability to methods that analyse loci formed of multiple linked polymorphic sites (e.g. Lohse et al., 2011) because the sequences generated are usually shorter than the distance between SNPs. RNAseq can involve more laborious and time consuming lab protocols compared to DNA library preparations, and extracting large enough quantities of good quality RNA is more difficult than the analogous processes for DNA (Ozsolak and Milos, 2011; Edwards et al., 2015).
Targeted enrichment involves selecting regions from the genome prior to sequencing (Mamanova et al., 2010). Originally it was primarily used in studies of human disease genes (e.g. Hodges et al., 2007a) but more recently it has been used to assess phylogenetic relationships in different taxa (Nicholls et al., 2015; Heyduk et al., 2016; Jones and Good, 2016). One advantage of targeted enrichment over RADseq is that the probes used for selecting genomic regions can be designed so that fragments overlap, allowing the assembly of larger contigs (continuous stretches of DNA built from overlapping sequence reads). This increases the chances of obtaining longer sequence blocks that contain multiple linked variable sites appropriate for gene tree inference methods (McCormack et al., 2013). The main drawback of such targeted enrichment methods is that the non-random approach to sampling the genome requires some prior knowledge of the genome to enable target probe design (Jones and Good, 2016).

In contrast, whole genome sequencing using short read technologies can be done using small starting quantities of DNA (< 50 ng) and using relatively straightforward library preparation protocols. It also enables the generation of datasets made up of large numbers of multiple loci (multilocus data). The difficulty in using whole genome short read data is the requirement for a reference genome to which the sequence reads can be aligned. If a reference genome is unavailable, an assembly will need to be built prior to further processing, which can be costly and computationally demanding (Jones and Good, 2016). One approach used in population genomic analyses, maximises the use of the sequence data generated per individual by combining the data across individuals from the same species to enable the assembly of a de novo reference sequence (Hearn et al., 2014). Assuming roughly equal coverage is achieved across individuals, this approach avoids reference bias (i.e. where reads preferentially map to the reference sequence because of the level of similarity between the sample and the individual(s) used to build the reference), an advantage when dealing with structured populations.

2.2.2 Study system and design
The east coast of Australia has several geological and topographical features that have been implicated as whole or partial barriers to gene flow in a range of species (see Chapter 1) (Moritz et al., 2000; Pope et al., 2001; Schäuble and Moritz, 2001; Garrick et al., 2004; Nicholls and Austin, 2005; Brown et al., 2006; Hodges et al., 2007b; Krosch et al., 2009; Moritz et al., 2009; Rossetto et al., 2009; Edwards and Melville, 2010; Beavis et al., 2011; Garrick et al., 2012; MacQueen et al., 2012; Bryant
Population genomics of a fig wasp community

and Fuller, 2014; Mellick et al., 2014; Bryant and Krosch, 2016). Using the geographic distribution of these divides as a guide, I identified three regions of potential interest from which to sample target fig wasp communities (Figure 2.1). The Burdekin Gap (BG) and St. Lawrence Gap (StLG) separate the two more northern sites. The BG is a dry woodland corridor in the eastern forest belt that stretches out to the coast to the south of Townsville in Queensland (Kelly et al., 2013) (Figure 1.3, Chapter 1). It is approximately 200 km wide and separates fig-containing forests to the north and south (Chapple et al., 2011b). The StLG is a smaller dry habitat corridor located between the towns of Mackay and Rockhampton (Chapple et al., 2011b) (Figure 1.3, Chapter 1). To explore the potential impact of these features on fig wasp community population structure, I sampled either side of these gaps, from Townsville up to Cairns to the north (henceforth the North population) and to the south of the gaps, predominantly around the Brisbane area (henceforth the South population). Samples were also collected from a third more southerly region around Sydney (henceforth the Sydney population). The South and Sydney populations are separated by the Hunter Valley (HV) (Figure 2.1), a dry lowland area just south of Newcastle (Figure 1.3, Chapter 1). The Brisbane Valley Barrier is a known barrier that is located within the South population (Figure 2.1). The valley is surrounded by mountainous regions south of Brisbane that have been suggested previously as an explanation for the presence of biogeographic divides between sibling fig wasp species between Queensland and New South Wales (Darwell, 2013).
I wanted to sample as many trophic levels as possible from the three populations described above from the figs of three species, *Ficus obliqua*, *Ficus rubiginosa* and *Ficus watkinsiana*. *Ficus watkinsiana* and *F. rubiginosa* are endemic to Australia whereas the native range of *F. obliqua* also extends to New Guinea and certain pacific islands (not including subsequent introductions) (Dixon et al., 2001; Lopez-Vaamonde et al., 2002; Dixon, 2003). All three species are distributed from northern Queensland to northern New South Wales; *F. obliqua* extends southwards to Sydney and *F. rubiginosa* is found as far south as the New South Wales/Victoria border (Figures 1.7, 1.9, 1.10, Chapter 1) (Dixon et al., 2001; Lopez-Vaamonde et al., 2002; Dixon, 2003). The aim of the study was to sample two individuals per population for each available species and generate low coverage genomic datasets for each. Across all three fig species, I generated genomic data for 11 wasp species, briefly described below. Table 2.1 details all the individuals sampled, their sampling locations and collection dates.
<table>
<thead>
<tr>
<th>Fig species</th>
<th>Population</th>
<th>Wasp genus</th>
<th>Samples sequenced^*</th>
<th>Sample locations</th>
<th>Sample coordinates</th>
<th>Date collected (and collector initials)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ficus rubiginosa</td>
<td>North</td>
<td>Pleistodentes</td>
<td>FRN_364_POL, FRN_J7_POL</td>
<td>Lake Tinaroo, QLD, Atherton, QLD</td>
<td>17.18° S, 145.57° E, 17.26° S, 145.48° E</td>
<td>22/08/2014 (LC), 01/12/2007 (JC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sycoscapter</td>
<td>FRN_350_SYC, FRN_274_SYC1, FRN_332_SYC</td>
<td>Castle Hill, Townsville, QLD, Mount Stuart, Townsville, QLD, Pallarenda, Townsville, QLD</td>
<td>19.25° S, 146.80° E, 19.35° S, 146.78° E, 19.20° S, 146.76° E</td>
<td>19/08/2014 (LC), 18/08/2014 (LC), 19/08/2014 (LC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Philotrypesis</td>
<td>FRN_310_PHIL, FRN_407_PHIL</td>
<td>Pallarenda, Townsville, QLD, Granite Gorge, QLD</td>
<td>19.20° S, 146.76° E, 17.04° S, 145.35° E</td>
<td>19/08/2014 (LC), 23/08/2014 (LC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Watshamiella</td>
<td>FRN_299_WAT, FRN_337_WAT</td>
<td>Pallarenda, Townsville, QLD, Pallarenda, Townsville, QLD</td>
<td>19.20° S, 146.76° E, 19.20° S, 146.76° E</td>
<td>19/08/2014 (LC), 19/08/2014 (LC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Herodotia</td>
<td>FRN_261_HERO, FRN_349_HERO, FRN_372_HERO</td>
<td>Pallarenda, Townsville, QLD, Castle Hill, Townsville, QLD, Herberton, QLD</td>
<td>19.20° S, 146.76° E, 19.25° S, 146.80° E, 17.38° S, 145.38° E</td>
<td>19/08/2014 (LC), 19/08/2014 (LC), 22/08/2014 (LC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eukobelea</td>
<td>FRN_310_EUKO, FRN_406_EUKO</td>
<td>Pallarenda, Townsville, QLD, Granite Gorge, QLD</td>
<td>19.20° S, 146.76° E, 17.04° S, 145.35° E</td>
<td>19/08/2014 (LC), 23/08/2014 (LC)</td>
</tr>
<tr>
<td></td>
<td>South</td>
<td>Pleistodentes</td>
<td>FRS_234_POL, FRS_511_POL</td>
<td>Broadbeach, Gold Coast, QLD, Nambucca Heads, NSW</td>
<td>28.03° S, 153.43° E, 30.64° S, 152.99° E</td>
<td>06/07/2014 (JC), 16/05/2013 (TS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sycoscapter</td>
<td>FRS_32_SYC, FRS_397_SYC, FRS_238_SYC, FRS_508_SYC</td>
<td>Noosa, QLD, Port Macquarie, NSW, Broadbeach, Gold Coast, QLD, Nambucca Heads, NSW</td>
<td>26.36° S, 152.96° E, 31.43° S, 152.90° E, 28.03° S, 153.43° E, 30.64° S, 152.99° E</td>
<td>07/11/2012 (JC), 14/05/2013 (TS), 04/07/2014 (JC), 16/05/2013 (TS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Philotrypesis</td>
<td>FRS_247_PHIL, FRS_110_PHIL, FRS_511_PHIL, FRS_40_PHIL</td>
<td>Main Beach, Gold Coast, QLD, Noosa, QLD, Nambucca Heads, NSW, Broadbeach, Gold Coast, QLD</td>
<td>27.96° S, 153.42° E, 26.36° S, 152.96° E, 30.64° S, 152.99° E, 28.03° S, 153.43° E</td>
<td>04/07/2014 (JC), 06/11/2012 (JC), 16/05/2013 (TS), 02/10/2013 (JC)</td>
</tr>
<tr>
<td>Wasp genus</td>
<td>Population</td>
<td>Sample locations</td>
<td>Sample coordinates</td>
<td>Date collected (and collector initials)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>------------------</td>
<td>--------------------</td>
<td>----------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRS 239_WAT</td>
<td>Main Beach, Gold Coast, QLD</td>
<td>27.96° S, 153.42° E</td>
<td>04/07/2014 (JC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRS 514_WAT</td>
<td>Nambucca Heads, NSW</td>
<td>30.64° S, 152.99° E</td>
<td>16/05/2013 (TS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRS 110_HERO</td>
<td>Noosa, QLD</td>
<td>26.36° S, 152.96° E</td>
<td>06/11/2012 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRS 239_HERO</td>
<td>Main Beach, Gold Coast, QLD</td>
<td>27.96° S, 153.42° E</td>
<td>04/07/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRS 517_HERO</td>
<td>Forster, NSW</td>
<td>27.96° S, 153.51° E</td>
<td>17/05/2013 (TS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRS 244_HERO</td>
<td>Main Beach, Gold Coast, QLD</td>
<td>27.96° S, 153.42° E</td>
<td>04/07/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRS 411_FUKO</td>
<td>Broadbeach, Gold Coast, QLD</td>
<td>28.03° S, 153.43° E</td>
<td>02/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRS 239_EUKO</td>
<td>Main Beach, Gold Coast, QLD</td>
<td>27.96° S, 153.42° E</td>
<td>04/07/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1085_POL</td>
<td>Wisemans Ferry, NSW</td>
<td>33.38° S, 150.98° E</td>
<td>16/12/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_2884_POL</td>
<td>Shelly Beach, Manly, NSW</td>
<td>33.80° S, 151.29° E</td>
<td>18/06/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_2873_SYC</td>
<td>Little Manly Cove, NSW</td>
<td>33.80° S, 151.29° E</td>
<td>18/06/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_2872_SYC</td>
<td>Shelly Beach, Manly, NSW</td>
<td>33.80° S, 151.29° E</td>
<td>18/06/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1964_HERO</td>
<td>Forster, NSW</td>
<td>33.82° S, 151.25° E</td>
<td>02/04/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1941_SYC</td>
<td>Helm Beach, NSW</td>
<td>33.38° S, 150.98° E</td>
<td>03/04/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1652_PHIL</td>
<td>Wisemans Ferry, NSW</td>
<td>33.38° S, 150.98° E</td>
<td>12/09/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukobelea</td>
<td>FRSYD_567_WAT</td>
<td>Little Manly Cove, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1705_WAT</td>
<td>Shelly Beach, Manly, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1822_PHIL</td>
<td>Wisemans Ferry, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1821_PHIL</td>
<td>Balmoral Beach, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1820_PHIL</td>
<td>Shelly Beach, Manly, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1819_PHIL</td>
<td>Wisemans Ferry, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1818_PHIL</td>
<td>Little Manly Cove, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>30/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1817_PHIL</td>
<td>Penrith, NSW</td>
<td>33.75° S, 150.70° E</td>
<td>03/04/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_1816_PHIL</td>
<td>Little Manly Cove, NSW</td>
<td>33.75° S, 150.70° E</td>
<td>03/04/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRSYD_381_HERO</td>
<td>Penrith, NSW</td>
<td>33.75° S, 150.70° E</td>
<td>17/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_380_HERO</td>
<td>Little Manly Cove, NSW</td>
<td>33.75° S, 150.70° E</td>
<td>17/10/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_382_HERO</td>
<td>Forster, NSW</td>
<td>33.80° S, 151.28° E</td>
<td>20/09/2013 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRSYD_383_HERO</td>
<td>Mareeba, QLD</td>
<td>17.00° S, 145.43° E</td>
<td>23/08/2014 (LC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FON_403_POL</td>
<td>Heberton, QLD</td>
<td>17.38° S, 145.38° E</td>
<td>22/08/2014 (LC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FON_398_POL</td>
<td>Mount Stuart, Townsville, QLD</td>
<td>19.35° S, 146.78° E</td>
<td>26/08/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FON_339_SYC</td>
<td>Mount Stuart, Townsville, QLD</td>
<td>19.35° S, 146.78° E</td>
<td>26/08/2014 (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FON_TrinBeach_SYC</td>
<td>Trinity Beach, QLD</td>
<td>16.78° S, 145.70° E</td>
<td>26/08/2014 (C)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig species**

Ficus oblivia
<table>
<thead>
<tr>
<th>Fig species</th>
<th>Population</th>
<th>Wasp genus</th>
<th>Samples sequenced^*</th>
<th>Sample locations</th>
<th>Sample coordinates</th>
<th>Date collected (and collector initials)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Philotrypesis</td>
<td>FON_340_PHIL, FON_PDoug_PHIL</td>
<td>Mount Stuart, Townsville, QLD, Port Douglas, QLD</td>
<td>19.35° S, 146.78° E, 16.48° S, 145.46° E</td>
<td>18/08/2014 (LC), 26/08/2014 (LC)</td>
</tr>
<tr>
<td></td>
<td>South</td>
<td>Pleistodontes</td>
<td>FOS_183_POL, FOS_187_POL</td>
<td>Broadbeach, Gold Coast, QLD, Woy Woy, NSW</td>
<td>28.03° S, 153.43° E, 33.49° S, 151.32° E</td>
<td>02/10/2013 (JC), 08/01/2014 (JC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sycoscapter</td>
<td>FOS_186_SYC, FOS_199_SYC</td>
<td>Woy Woy, NSW</td>
<td>33.49° S, 151.32° E, 33.49° S, 151.32° E</td>
<td>08/01/2014 (JC), 25/03/2014 (JC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Philotrypesis</td>
<td>FOS_186_PHIL, FOS_188_PHIL</td>
<td>Woy Woy, NSW</td>
<td>33.49° S, 151.32° E, 33.49° S, 151.32° E</td>
<td>08/01/2014 (JC), 08/01/2014 (JC)</td>
</tr>
<tr>
<td>Ficus watkinsiana</td>
<td>North</td>
<td>Pleistodontes</td>
<td>FWN_01-8_POL, FWN_161_POL</td>
<td>Kairi, QLD, Kamerunga, QLD</td>
<td>17.21° S, 145.55° E, 16.87° S, 145.68° E</td>
<td>17/01/2001 (JC), 09/08/2009 (JC)</td>
</tr>
<tr>
<td></td>
<td>South</td>
<td>Pleistodontes</td>
<td>FWS_03-250_POL, FWS_137_POL</td>
<td>Settlers Rise, QLD, Main Range, QLD</td>
<td>27.68° S, 153.26° E, 28.07° S, 152.41° E</td>
<td>15/01/2004 (JC), 29/11/2008 (JC)</td>
</tr>
</tbody>
</table>

**Table 2.1**: Details of all the sequenced individuals, sampling locations, sampling date and initials of the sample collector. ^*The sequenced samples were all labelled according to the tree of origin (e.g. FR for *Ficus rubiginosa*), the population (e.g. N (North)), the tube code (e.g. 183) and the genus of wasp (e.g. SYC for *Sycoscapter*, in the case of *Pleistodontes*, POL). † This individual failed to sequence. Collector initials – LC (Lisa Cooper), JC (James Cook), TS (Tim Sutton).
The most extensively studied species of the three figs is *F. rubiginosa*, which produces mid-sized figs (Figure 1.8, Chapter 1) that are easy to collect and show a diverse range of wasp inhabitants (Segar et al., 2014). *Ficus rubiginosa* is associated with a complex of five sister species of its pollinator *Pleistodontes imperialis* (Darwell et al., 2014), currently known as *P. imperialis* species 1-5 and found along the east coast with differing distributions (Figure 1.11, Chapter 1). Species 1 is the only species found in all three populations (Darwell et al., 2014), so it was chosen as the representative *F. rubiginosa* pollinator. Parasitoids of the pollinators, from the genus *Sycoscapter*, are present as two morphologically distinguishable (in females) forms based on ovipositor length (long and short) (Segar, 2011). Both forms of *Sycoscapter* parasitoids were sampled, however, only the long morph was sampled from all three populations. It only made sense to sample the short morph from Queensland (one ‘population’) and New South Wales (other ‘population’) given its distribution (Figure 2.2) (Note: this sampling does not fit in strictly with the designation of the North, South, Sydney populations outlined above). Alongside the pollinator and its parasitoids, the inquiline genus *Philotryptesis*, and the hyper-parasitoid genus *Watshamiella* were sampled from all three populations. Due to the ease of sampling and availability of previous collections from *F. rubiginosa* I was also able to sample two galler species of the genera *Herodotia* and *Eukobelea*. *Herodotia* was again sampled from all three populations but due to the distribution of *Eukobelea* (Figure 2.3), I only collected *Eukobelea* samples from the North and South sites.
Figure 2.2: Map of the east coast of Australia showing the distribution of the two *Sycoscapter* short species (parasitoids). It shows that the two species overlap around Brisbane. Figure adapted from one in Darwell (2013).

Figure 2.3: Map of the east coast of Australia showing the distribution of the two *Eukobelea* species (gallers). Figure adapted from one in Darwell (2013).
*Ficus watkinsiana* has the largest figs of the three species (Figure 1.8, Chapter 1) but due to the tree’s height (up to 50 m) and growth locations it is much harder to sample than the other two species. As a result, I was only able to sample its pollinator, *Pleistodontes nigriventris* (Lopez-Vaamonde et al., 2002) from the North and South populations.

*F. obliqua* is the smallest of the figs sampled (Figure 1.8, Chapter 1) and correspondingly its wasp inhabitants are also some of the smallest I collected. It is more difficult to obtain samples of the larger NPFW species as they are rarely found in these small figs so I could only sample its *Pleistodontes* pollinator, the parasitoid *Sycoscapter* and the inquiline *Philotrypesis*, all from the North and South populations only (Note: some of the *F. obliqua* South samples are actually from Sydney due to a lack of available specimens from the Brisbane area).

### 2.2.3 Haplodiploidy

All taxa investigated in this study are insects of the order Hymenoptera, which exhibit a phenomenon known as haplodiploidy, in which fertilised eggs form diploid females and unfertilised eggs develop into haploid males (Normark, 2004; Li et al., 2016). This is highly advantageous when assessing genetic diversity using sequence-based analyses as males possess only a single allele at each locus (Xiao et al., 2013). The lack of heterozygous sites makes the variant calling process a lot simpler (Hearn et al., 2014). It also avoids the need to correctly identify which haploid genome the two alleles that are present in diploids come from (a process termed ‘phasing’). To be confident when calling heterozygous sites in a diploid organism, each individual needs to be sequenced at a higher coverage (the number of times a particular part of the genome is sequenced) compared to a haploid organism. This is because it is expected that approximately half the reads covering a truly heterozygous site will show one allele with the other half showing the alternate allele. In order to see this pattern (and to distinguish it from varying frequencies of alternate alleles resulting from sequencing error) enough reads must cover each site to be confident in these calls. In a haploid organism however if two alleles are called for a single site it must represent an error (Hearn et al., 2014). For a given sequencing budget this allows the number of individuals/taxa to be prioritised over the requirement of greater coverage per individual.
2.2.4 Sample selection

The small sample numbers per population used in this study, and the large investment of resources and time in each of them, meant it was important to attempt to accurately identify each individual prior to whole genome sequencing. Species identification was a challenge for several reasons: 1) the small average size of the wasps (Figure 2.4), 2) the desire to sequence males (males are generally even smaller than the females (Figure 2.4) and can be morphologically similar across species (Cook et al., 2015)), and 3) the number of morphologically cryptic sister species shown to be present in several of the sampled genera (Segar, 2011; Darwell, 2013). Male fig wasps are generally harder to collect and therefore not as common in collections as females. Males do not normally exit the figs (Xiao et al., 2013; Sutton et al., 2016) and so each fig must be dissected to extract them. Ideally I wanted each individual per population to be sampled from a different site to increase the chance of sampling the coalescent variation present within the population. If this was not possible it was essential to avoid sampling two individuals from the same fig due to the high levels of sib-mating in some of the fig wasp species (Greeff et al., 2009; Sutton et al., 2016). Given the small numbers of available candidate specimens for some taxa, the requirement of different sites was easier to meet for some species over others. DNA sequence barcoding was used in an attempt to check whether identified morpho-taxa also belonged to the same molecular species (strictly, MOTUs, molecular operational taxonomic units). This process uses single locus sequencing, usually of an mtDNA locus, to define genetic groups at various taxonomic levels (Hebert et al., 2003). It relies on there being a ‘barcoding gap’ between inter- and intra-specific variation. The technique has been used successfully to establish interspecies relationships but can encounter difficulties when used to assess groupings in closely related taxa (Hurst and Jiggins, 2005). For example, there may be overlap between inter- and intra-specific variation preventing clear delimitation of groups (e.g. Wiemers and Fiedler, 2007), or the mitochondria may have been exposed to selective sweeps caused by co-infection of the wasp host with maternally inherited bacterial endosymbionts such as Wolbachia (Hurst and Jiggins, 2005). Given the challenges outlined here, the path to sample selection can be broken down into three main points of consideration: 1) Selection of a ‘wish list’ of taxa based on biological interest, 2) Identification of putative specimens, with identification confirmed morphologically and/or by molecular barcoding, 3) Of those species with enough identified specimens for useful sampling, which produced enough DNA for library preparation, and of those, which produced good enough library DNA fragment size distributions for sequencing.
Figure 2.4: Examples of the sampled fig wasps sequenced in this study. Photographs were taken down the microscope. The upper left photo shows a winged female and wingless male pollinator (Pleistodontes ex F. rubiginosa). The lower left photo shows two Sycoscapter (ex F. rubiginosa) males. The photos on the right show a male Eukobelea (ex F. rubiginosa) wasp next to a 6mm staple to demonstrate their tiny size.

2.2.5 Objectives

The overall aim of this chapter is to generate low coverage genomic datasets at the individual level for a community of fig wasps distributed across the east coast of Australia. The chapter addresses the following questions related to the generation of such genomic datasets: Can morphologically cryptic male fig wasps be correctly identified using a fragment of the mitochondrial cytochrome b (cytb) gene as a molecular barcode? Can a high enough quantity of DNA be extracted from each individual wasp to successfully prepare high throughput libraries for sequencing on the Illumina HiSeq? What level of contamination is present in the assemblies and, if present, what are the main contributors? Is the level of sequencing coverage achieved equal across species? Is the meta-assembly approach, i.e. using multiple individuals across populations per species to build a reference assembly, effective and does the quality of these assemblies effect diversity measures? Below I outline my hypotheses in response to these questions and briefly state the reasons for them.
Population genomics of a fig wasp community

The cytb barcode fragment has been used previously to identify the target species of this study, therefore I expect it to successfully delimit the relationships of the samples tested. The results from the single locus mtDNA barcodes will be confirmed using concatenated phylogenies based on average genetic distances calculated within and between populations from genome-wide SNP calls. As far as I am aware no high throughput libraries from individual fig wasps have been made prior to this study. Consequently, I have no a priori expectation in terms of the DNA quantities that could be achieved. However, the Nextera kits used in this study have been used by other members of the lab group to successfully prepare and sequence libraries from other chalcids indicating it may be possible to achieve this for the wasp samples used here. Fig wasps are known to harbour diverse microbial communities including the endosymbiont, Wolbachia. For this reason, I hypothesise I am likely to find contaminant reads (i.e. reads not originating from fig wasp tissue) present in the assemblies. The specimens were collected to minimise the presence of any plant material being present on the outside of the wasps and as adult males are not known to feed on the figs I expect to see little contamination from plant material in the assemblies. The pooling of the libraries prior to sequencing was carried out with the aim of achieving equal sequencing depth across species. However, due to technical difficulties in estimating the average fragment size of each library and because the calculations use the same estimate of genome size (based on the only published fig wasp genome) across all the species sampled, I expect to find a certain level of fluctuation in the sequencing coverage achieved. Assuming the samples have been identified correctly I anticipate the meta-assembly approach to be an effective method for producing a reference assembly from multiple low coverage datasets. The resultant assemblies however will likely be highly fragmented as a consequence of the higher level of variation present between individuals from multiple populations. This approach has been used successfully in previous studies and although it mitigates the issue of reference bias, it is possible the approach will bias toward coding regions as these tend to be more conserved and therefore can assemble better. To establish whether this is likely to be a problem in the datasets generated here I plan to assess whether there is any correlation present between assembly quality and genetic diversity.
2.3 Methods

2.3.1 Sample collection

Most sampling was conducted between November 2012 and August 2014 with a few additional samples obtained between 2001 and 2009. Table 2.1 details the collection dates, identity of the collector and locations of the samples used for the HiSeq sequencing. The sampling regions are grouped into North, South and Sydney populations.

Near-to-ripe figs were collected and placed individually into specimen pots, with the exception of the small *F. obliqua* figs, which were pooled into groups from the same tree. Once the wasps started to emerge (12-24 hours after collection depending on fig ripeness) the figs were dissected and live males were placed directly into 70% ethanol to preserve for DNA extraction. Due to potentially high levels of sib-mating (Greeff et al., 2009; Sutton et al., 2016) in fig wasps, no individuals of the same species collected from the same fig were used in this study.

2.3.2 Sample DNA barcoding

There are a number of cryptic species present in fig wasp genera making morphological identification nearly impossible for some species (Darwell, 2013), especially in males. For this reason, I DNA barcoded all samples prior to using them in high throughput sequencing library preparation. Samples for barcoding were selected based on genera level morphological identification and location, ideally to select at least 4 individuals per genus and population, and within each population at least 2 individuals from different sites. DNA was extracted from whole male wasps using the Qiagen DNeasy Blood and Tissue Extraction kit. The protocol was adapted slightly to maximise DNA yield from these extremely small wasps (Figure 2.4). The Purification of Total DNA from Animal Tissues (Spin-Column) Protocol was followed with the following adaptions. Step 1: The individual wasps were placed in 180 µl of buffer ATL and crushed using a mini-pestle. Step 2: Riboshredder RNase from Epicentre was used in place of RNase A. Steps 7-8: Buffer EB was used in place of Buffer AE for the elutions as Buffer AE contains EDTA which will interfere with the downstream library preparation protocol. The elutions were done using smaller volumes (25 µl) than recommended and the samples incubated for longer (5 minutes) (protocol adaptions optimised by J. Nicholls). All the samples were sequenced for a 433 base pair fragment of the mitochondrial cytochrome b (cytb) gene using the CB1/CB2 primers (Jermiin and Crozier, 1994). These primers failed to amplify for some of the individuals of the genera.
Population genomics of a fig wasp community

*Philotrypesis, Watshamiella, Herodotia and Eukobelea.* In these cases another primer pair (CP1/CP2) (Harry et al., 1998), which amplifies a longer (760 bp) but overlapping fragment of the cytb gene, were used, either together or by using the CB1/CP2 combination. Table 2.2 details the samples barcoded and the primer pairs used. To minimise the amount of DNA lost to the barcoding process only 0.3 µl DNA was used per PCR reaction. The remainder of the PCR mix consisted of 2 µl BSA (10 mg/ml), 2 µl 10X PCR buffer, 0.8 µl MgCl₂ (50 mM), 0.3 µl of each primer (20 µM), 0.16 µl dNTPs (each 25 mM) and 0.1 µl Taq (Bioline 5U/µl), made up to 20 µl with autoclaved MilliQ water. Amplification of all fragments was carried out using a Bio-Rad S1000 thermal cycler for 2 minutes at 94°C, 35 cycles of 30 seconds at 94°C, 30 seconds at 48°C, 40 seconds at 72°C, and a final elongation step of 5 minutes at 72°C. The PCR products were visualised using a 2% agarose gel and cleaned using a shrimp alkaline phosphatase and exonuclease 1 protocol. 2.5 µl of SAPExo1 mix (1.425 µl SAP dilution buffer, 1 µl SAP (1U), 0.075 µl Exo1 (1.5U)) was added to each sample before being run on a Bio-Rad S1000 thermal cycler for 40 minutes at 37°C followed by 15 minutes at 94°C. The quantity of PCR product used in the subsequent sequencing reaction was determined from the brightness of the band on the gel. Only the forward strand was sequenced for each individual. The fragments were sequenced using BigDye chemistry on an ABI 3730 machine at the Edinburgh Genomics sequencing facility in Edinburgh. The sequences have been deposited on GenBank (accession numbers: MF597796-MF597932).
<table>
<thead>
<tr>
<th>Fig species</th>
<th>Population</th>
<th>Wasp genus</th>
<th>No. collection sites</th>
<th>No. individuals available*</th>
<th>No. individuals barcoded**</th>
<th>No. libraries made (used)</th>
<th>Samples sequenced^*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ficus rubiginosa</td>
<td>North</td>
<td>Pleistodones 9</td>
<td>295 (44 (5), 153 (21), 17 (1), 59 (6), 19 (5), 17 (2), 13 (2), 1 (1), 2 (2))</td>
<td>11 (CB1/CB2)</td>
<td>2 (2)</td>
<td>FRN_364_POL FRN_J7_POL</td>
<td></td>
</tr>
<tr>
<td>Sycoscapt er 7</td>
<td>46 (7 (1), 17 (10), 1 (1), 2 (1), 1 (1), 14 (1), 4 (2))</td>
<td>19 (CB1/CB2)</td>
<td>3 (long) (2)</td>
<td>FRN_350_SYC FRN_274_SYC1 FRN_332_SYC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philotrypesis 5</td>
<td>67 (40 (9), 1 (1), 14 (5), 7 (3), 5 (1))</td>
<td>9 (CB1/CB2)</td>
<td>2 (short) (1)</td>
<td>FRN_310_PHIL FRN_407_PHIL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watshamiella 2</td>
<td>7 (6 (6), 1 (1))</td>
<td>5 (CB1/CB2)</td>
<td>3 (CB1/CB2)</td>
<td>FRN_299_WAT FRN_337_WAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herodotia 3</td>
<td>13 (10 (7), 1 (1), 2 (2))</td>
<td>3 (CB1/CB2)</td>
<td>3 (CB1/CB2)</td>
<td>FRN_261_HERO FRN_349_HERO FRN_372_HERO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukobelea 4</td>
<td>6 (2 (1), 1 (1), 2 (1), 1 (1))</td>
<td>4 (CB1/CP2)</td>
<td>4 (CB1/CP2)</td>
<td>FRN_310_EUKO FRN_406_EUKO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pleistodones South</td>
<td>6</td>
<td>91 (5 (1), 48 (8), 19 (4), 2 (2), 14 (5), 3 (2))</td>
<td>6 (CB1/CB2)</td>
<td>2 (2)</td>
<td>FRS_234_POL FRS_511_POL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sycoscapt er 7</td>
<td>60 (5 (5), 25 (8), 19 (10), 1 (1), 1(1), 5 (4), 4 (2))</td>
<td>7 (CB1/CB2)</td>
<td>3 (long) (2)</td>
<td>FRS_32_SYC FRS_397_SYC FRS_238_SYC FRS_508_SYC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philotrypesis 4</td>
<td>53 (14 (4), 15 (7), 15 (4), 9 (3))</td>
<td>6 (CB1/CB2)</td>
<td>5 (CB1/CB2)</td>
<td>FRS_247_PHIL FRS_110_PHIL FRS_511_PHIL FRS_40_PHIL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fig species</td>
<td>Population</td>
<td>Wasp genus</td>
<td>No. collection sites</td>
<td>No. individuals available*</td>
<td>No. individuals barcoded**</td>
<td>No. libraries made (used)</td>
<td>Samples sequenced^*</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>------------</td>
<td>----------------------</td>
<td>----------------------------</td>
<td>----------------------------</td>
<td>--------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td><em>Watshamiella</em></td>
<td>6</td>
<td><em>Watshamiella</em></td>
<td>16 (4 (2), 2 (2), 5 (3), 1 (1), 3 (2), 1 (1))</td>
<td>6 (CB1/CB2)</td>
<td>3 (2)</td>
<td>FRS_239_WAT, FRS_514_WAT</td>
<td></td>
</tr>
<tr>
<td><em>Herodotia</em></td>
<td>3</td>
<td><em>Herodotia</em></td>
<td>8 (2 (2), 2 (2), 4 (1))</td>
<td>5 (CB1/CB2)</td>
<td>4 (4)</td>
<td>FRS_110_HERO, FRS_239_HERO, FRS_517_HERO, FRS_244_HERO</td>
<td></td>
</tr>
<tr>
<td><em>Eukobelea</em></td>
<td>3</td>
<td><em>Eukobelea</em></td>
<td>6 (4 (3), 1 (1), 1 (1))</td>
<td>3 (CB1/CB2 and CB1/CP2)</td>
<td>2 (2)</td>
<td>FRS_41_EUKO, FRS_239_EUKO</td>
<td></td>
</tr>
<tr>
<td><em>Pleistodontes</em></td>
<td>6</td>
<td><em>Pleistodontes</em></td>
<td>12 (2 (1), 2 (1), 2 (1), 2 (1), 2 (1), 2 (1))</td>
<td>4 (CB1/CB2)</td>
<td>4 (2)</td>
<td>FRSYD_1065_POL, FRSYD_2884_POL</td>
<td></td>
</tr>
<tr>
<td><em>Sycoscapter</em></td>
<td>6</td>
<td><em>Sycoscapter</em></td>
<td>15 (3 (2), 2 (1), 2 (1), 2 (1), 4 (2), 2 (1))</td>
<td>8 (CB1/CB2)</td>
<td>2 (long) (2)</td>
<td>FRSYD_89_SYC, FRSYD_2873_SYC, FRSYD_1941_SYC</td>
<td></td>
</tr>
<tr>
<td><em>Philotrypesis</em></td>
<td>5</td>
<td><em>Philotrypesis</em></td>
<td>10 (2 (1), 2 (2), 4 (2), 1 (1), 1 (1))</td>
<td>4 (CB1/CB2, CB1/CP2 and CP1/CP2)</td>
<td>4 (2)</td>
<td>FRSYD_1822_PHIL, FRSYD_165_PHIL</td>
<td></td>
</tr>
<tr>
<td><em>Watshamiella</em></td>
<td>5</td>
<td><em>Watshamiella</em></td>
<td>14 (4 (2), 2 (1), 3 (1), 2 (1), 2 (2))</td>
<td>6 (CB1/CB2, CB1/CP2 and CB1/CP2)</td>
<td>4 (2)</td>
<td>FRSYD_567_WAT, FRSYD_1905_WAT1</td>
<td></td>
</tr>
<tr>
<td><em>Herodotia</em></td>
<td>4</td>
<td><em>Herodotia</em></td>
<td>8 (2 (1), 2 (1), 2 (1), 2 (1))</td>
<td>4 (CB1/CB2 and CB1/CP2)</td>
<td>3 (3)</td>
<td>FRSYD_381_HERO, FRSYD_293_HERO, FRSYD_1964_HERO</td>
<td></td>
</tr>
<tr>
<td><em>Ficus obliqua</em></td>
<td>North</td>
<td><em>Pleistodontes</em></td>
<td>99 (46 (8), 18 (2), 2 (1), 29 (3), 1 (1), 3 (1))</td>
<td>6 (CB1/CB2)</td>
<td>4 (2)</td>
<td>FON_403_POL, FON_398_POL</td>
<td></td>
</tr>
<tr>
<td>Fig species</td>
<td>Population</td>
<td>Wasp genus</td>
<td>No. collection sites</td>
<td>No. individuals available*</td>
<td>No. individuals barcoded**</td>
<td>No. libraries made (used)</td>
<td>Samples sequenced^*</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>------------</td>
<td>---------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
<td>-------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><em>Sycoscapter</em></td>
<td>South</td>
<td>6</td>
<td>100 (11 (5), 37 (2), 3 (1), 32 (3), 15 (1), 2 (1))</td>
<td>6 (CB1/CB2)</td>
<td>4 (2)</td>
<td>FON_339_SYC, FON_TrinBeach_SYC</td>
<td></td>
</tr>
<tr>
<td><em>Philotrypesis</em></td>
<td>South</td>
<td>5</td>
<td>24 (3 (1), 5 (2), 12 (3), 2 (1), 2 (1))</td>
<td>5 (CB1/CB2)</td>
<td>2 (2)</td>
<td>FON_340_PHIL, FON_PDoug_PHIL</td>
<td></td>
</tr>
<tr>
<td><em>Watshamiella</em></td>
<td>South</td>
<td>2</td>
<td>4 (3 (2), 1 (1))</td>
<td>2 (CB1/CB2)</td>
<td>0</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><em>Pleistodentes</em></td>
<td>South</td>
<td>2</td>
<td>26 (20 (5), 6 (1))</td>
<td>4 (CB1/CB2)</td>
<td>4 (2)</td>
<td>FOS_183_POL1, FOS_187_POL</td>
<td></td>
</tr>
<tr>
<td><em>Sycoscapter</em></td>
<td>South</td>
<td>1</td>
<td>30 (4)</td>
<td>4 (CB1/CB2)</td>
<td>2 (2)</td>
<td>FOS_186_SYC, FOS_199_SYC</td>
<td></td>
</tr>
<tr>
<td><em>Philotrypesis</em></td>
<td>South</td>
<td>1</td>
<td>9 (3)</td>
<td>6 (CB1/CB2)</td>
<td>4 (2)</td>
<td>FOS_186_PHIL, FOS_188_PHIL1</td>
<td></td>
</tr>
<tr>
<td><em>Watshamiella</em></td>
<td>South</td>
<td>1</td>
<td>2 (2)</td>
<td>2 (CB1/CB2)</td>
<td>0</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><em>Ficus watkinsiana</em></td>
<td>North</td>
<td><em>Pleistodentes</em></td>
<td>3</td>
<td>40 (28 (2), 3 (1), 9 (1))</td>
<td>4 (CB1/CB2 and CB1/CP2)</td>
<td>3 (2)</td>
<td>FWN_01-8_POL, FWN_161_POL</td>
</tr>
<tr>
<td><em>Pleistodentes</em></td>
<td>South</td>
<td>4</td>
<td>32 (3 (1), 3 (1), 14 (2), 12 (1))</td>
<td>4 (CB1/CB2 and CB1/CP2)</td>
<td>2 (2)</td>
<td>FWS_03-250_POL, FWS_137_POL</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Details of the samples collected, barcoded, the number of libraries prepared and the samples sequenced. *Number individuals available written as total number of individuals per species per population (bold) then broken down (in brackets) to describe how many came from each site and then further brackets show how many figs/tubes were available from within that site. Where only a single site was sampled the number in brackets shows how many figs/tubes were available from within that site. **Number of individuals barcoded and the primer pairs used in the PCR reactions (the forward primer was always used for the sequencing reactions). ^*The sequenced samples were all labelled according to the tree of origin (e.g. FR for *Ficus rubiginosa*), the population (e.g. N (North)), the tube code (e.g. 183) and the genus of wasp (e.g. SYC for *Sycoscapter*, in the case of *Pleistodentes*, POL). *This individual failed to sequence.
The resultant chromatograms were inspected by eye and edited using MEGA version 6.06 (Tamura et al., 2013). Alignments were produced in MEGA using the aligner Muscle (Edgar, 2004) and exported as NEXUS files. I added a number of additional fig wasp sequences to the alignments to assist with the identification, provided by C. Darwell (Okinawa Institute of Science and Technology), J. Cook, T. Sutton and A. Montagu (Western Sydney University). The NEXUS files were converted to PHYLIP format using the sequence converter at http://sequenceconversion.bugaco.com/converter/biology/sequences/. The alignments were used to produce maximum likelihood (ML) trees in PhyML 3.0 (Guindon et al., 2010). The HKY85 model of substitution was used for each dataset. Support values were estimated using the aBayes method (Anisimova et al., 2011) implemented in PhyML 3.0. Values ≥ 0.95 were considered as strong support, values < 0.95 and ≥ 0.7 were considered as moderate support and those < 0.7 were considered as poor support. The trees were rooted using sequences provided by C. Darwell and J. Cook. Table 2.3 details the outgroup genera used for each dataset. The trees were edited and exported using FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) and further edited using Inkscape 0.91 (https://inkscape.org/en/).
Table 2.3: Details of the individuals used to root each of the mtDNA phylogenies.

<table>
<thead>
<tr>
<th>mtDNA dataset (wasp genus and subfamily)</th>
<th>Outgroup taxa used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleistodontes (Agaoninae)</td>
<td>P. blandus (ex F. glandifera) individual from New Guinea</td>
</tr>
<tr>
<td>Sycoscopter (Sycoryctinae)</td>
<td>Philotrypes (ex F. rubiginosa and F. obliqua) individuals</td>
</tr>
<tr>
<td>Philotrypes (Sycoryctinae)</td>
<td>Watshamiella (ex F. rubiginosa and F. obliqua) individuals</td>
</tr>
<tr>
<td>Watshamiella (Sycoryctinae)</td>
<td>Sycoscopter (ex F. rubiginosa and F. obliqua) individuals</td>
</tr>
<tr>
<td>Herodotia (Epichrysomallinae)</td>
<td>Meselatus (ex F. rubiginosa) – a large galler from subfamily Epichrysomallinae</td>
</tr>
<tr>
<td>Eukobelea (Sycophaginae)</td>
<td>Pseudidarnes (ex F. rubiginosa) – a large galler from subfamily Sycophaginae</td>
</tr>
</tbody>
</table>

2.3.3 High-throughput library preparation and sequencing

The high throughput sequencing libraries were made using Illumina Nextera DNA Sample Preparation Kits. The protocol requires 50 ng DNA (in a working volume of 20 µl) so the DNA extractions were quantified using a Qubit fluorometer (Life Technologies) to prepare them for use in the protocol. The samples that did not contain enough DNA in the elution volume (~33 µl) were dried on a heat block at 60°C and resuspended in the appropriate volume of EB Buffer. The samples and their quantities are shown in Table 2.4. The Nextera protocol was used following the manufacturers instructions apart from the volume of Tagmentation enzyme used in Stage 1, which was varied according to DNA starting quantity (Table 2.4).

Fragments for each individual were end-labelled using a unique pair of indices enabling each to be identified from a combined pool of several individuals. This allows these pools of multiple individuals to be run on a single lane of the sequencer. The library fragment size distribution was visualised by running 1 µl of each library on a high sensitivity DNA Bioanalyzer chip using the Agilent 2100 Bioanalyzer and following the manufacturers instructions. The traces from these
runs can be seen in Appendix A. Ideally the distributions would show a single peak around the predominant fragment size present in the library (for example see Figure 2.5). It is important for the predominant fragment size to be longer than the size of the reads sequenced e.g., if fragments are sequenced using 150 bp paired end reads, any fragments <150 bp will cause the sequencer to run out of fragment to sequence, ultimately wasting resources and nullifying any benefit gained from using paired end reads. This will also result in the de novo assembly of the reads being severely fragmented as the pairs will only ever cover very small fragments. If the quality of the trace was too poor to use (for example see Figure 2.5) I made a replacement library from a different individual where available. The samples were pooled as detailed in Appendix B prior to being run on the Illumina HiSeq at the Edinburgh Genomics facility in Edinburgh. The pooling volumes were calculated by aiming for ~6 fold coverage per individual assuming each lane would give 50 gigabases (Gb) of data. Two lanes of 125 bp paired-end reads were run to sequence all samples, with the exception of the F. watkinsiana pollinators. These were run on a lane of 150 bp paired-end reads. The short read data have been deposited at the ENA short read archive (in progress at time of printing).
<table>
<thead>
<tr>
<th>Sample</th>
<th>Estimated DNA amount (ng)</th>
<th>Tagmentation enzyme volume used (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FON_398_POL</td>
<td>18.28</td>
<td>1.8</td>
</tr>
<tr>
<td>FON_403_POL</td>
<td>21.13</td>
<td>2.1</td>
</tr>
<tr>
<td>FOS_183_POL1</td>
<td>23.75</td>
<td>2.3</td>
</tr>
<tr>
<td>FOS_187_POL</td>
<td>22.84</td>
<td>2.3</td>
</tr>
<tr>
<td>FON_339_SYC</td>
<td>25.68</td>
<td>2.6</td>
</tr>
<tr>
<td>FON_TrinBeach_SYC</td>
<td>27.70</td>
<td>2.8</td>
</tr>
<tr>
<td>FOS_186_SYC</td>
<td>48.26</td>
<td>4.8</td>
</tr>
<tr>
<td>FOS_199_SYC</td>
<td>40.50</td>
<td>4.1</td>
</tr>
<tr>
<td>FON_340_PHIL</td>
<td>28.79</td>
<td>2.9</td>
</tr>
<tr>
<td>FON_PDoug_PHIL</td>
<td>22.00</td>
<td>2.2</td>
</tr>
<tr>
<td>FOS_186_PHIL</td>
<td>36.00</td>
<td>3.6</td>
</tr>
<tr>
<td>FOS_188_PHIL1</td>
<td>31.35</td>
<td>3.0</td>
</tr>
<tr>
<td>FRN_364_POL</td>
<td>45.88</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_J7_POL</td>
<td>66.12</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_234_POL</td>
<td>37.98</td>
<td>4.0</td>
</tr>
<tr>
<td>FRS_511_POL</td>
<td>45.98</td>
<td>5.0</td>
</tr>
<tr>
<td>FRSYD_1065_POL</td>
<td>32.72</td>
<td>3.2</td>
</tr>
<tr>
<td>FRSYD_2884_POL</td>
<td>34.28</td>
<td>3.4</td>
</tr>
<tr>
<td>FRN_350_SYC</td>
<td>24.55</td>
<td>2.5</td>
</tr>
<tr>
<td>FRN_274_SYC1</td>
<td>93.48</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_332_SYC</td>
<td>30.49</td>
<td>3.0</td>
</tr>
<tr>
<td>FRS_32_SYC</td>
<td>27.72</td>
<td>3.0</td>
</tr>
<tr>
<td>FRS_397_SYC</td>
<td>46.17</td>
<td>4.6</td>
</tr>
<tr>
<td>FRS_238_SYC</td>
<td>45.79</td>
<td>4.5</td>
</tr>
<tr>
<td>FRS_508_SYC</td>
<td>52.20</td>
<td>5.0</td>
</tr>
<tr>
<td>FRSYD_1941_SYC</td>
<td>39.14</td>
<td>3.9</td>
</tr>
<tr>
<td>FRSYD_89_SYC</td>
<td>31.50</td>
<td>3.1</td>
</tr>
<tr>
<td>FRSYD_2873_SYC</td>
<td>31.50</td>
<td>3.1</td>
</tr>
<tr>
<td>FRN_310_PHIL</td>
<td>58.71</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_407_PHIL</td>
<td>58.71</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_110_PHIL</td>
<td>41.80</td>
<td>4.2</td>
</tr>
<tr>
<td>FRS_247_PHIL</td>
<td>53.84</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_511_PHIL</td>
<td>53.01</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_40_PHIL</td>
<td>41.04</td>
<td>4.0</td>
</tr>
<tr>
<td>FRSYD_1822_PHIL</td>
<td>29.45</td>
<td>2.9</td>
</tr>
<tr>
<td>FRSYD_165_PHIL</td>
<td>28.01</td>
<td>2.8</td>
</tr>
<tr>
<td>FRN_299_WAT</td>
<td>48.83</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_337_WAT</td>
<td>49.59</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_239_WAT</td>
<td>50.16</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_514_WAT</td>
<td>55.87</td>
<td>5.0</td>
</tr>
<tr>
<td>FRSYD_567_WAT</td>
<td>37.63</td>
<td>3.7</td>
</tr>
</tbody>
</table>
### Table 2.4

<table>
<thead>
<tr>
<th>Sample</th>
<th>Estimated DNA amount (ng)</th>
<th>Tagmentation enzyme volume used (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRSYD_1905_WAT1</td>
<td>50.14</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_261_HERO</td>
<td>404.70</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_349_HERO</td>
<td>133.92</td>
<td>5.0</td>
</tr>
<tr>
<td>FRN_372_HERO</td>
<td>167.96</td>
<td>4.5</td>
</tr>
<tr>
<td>FRS_110_HERO</td>
<td>486.00</td>
<td>4.3</td>
</tr>
<tr>
<td>FRS_239_HERO</td>
<td>222.12</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_517_HERO</td>
<td>255.96</td>
<td>5.0</td>
</tr>
<tr>
<td>FRS_244_HERO</td>
<td>283.10</td>
<td>4.6</td>
</tr>
<tr>
<td>FRSYD_381_HERO</td>
<td>236.74</td>
<td>5.0</td>
</tr>
<tr>
<td>FRNS_293_HERO</td>
<td>169.48</td>
<td>4.9</td>
</tr>
<tr>
<td>FRSYD_1964_HERO</td>
<td>315.40</td>
<td>4.6</td>
</tr>
<tr>
<td>FRN_310_EUKO</td>
<td>24.52</td>
<td>2.5</td>
</tr>
<tr>
<td>FRN_406_EUKO</td>
<td>20.52</td>
<td>2.0</td>
</tr>
<tr>
<td>FRS_41_EUKO</td>
<td>22.46</td>
<td>2.3</td>
</tr>
<tr>
<td>FRS_239_EUKO</td>
<td>27.11</td>
<td>2.7</td>
</tr>
<tr>
<td>FWN_01_8_POL</td>
<td>47.77</td>
<td>5.0</td>
</tr>
<tr>
<td>FWN_161_POL</td>
<td>58.25</td>
<td>5.0</td>
</tr>
<tr>
<td>FWS_03_250_POL</td>
<td>42.50</td>
<td>5.0</td>
</tr>
<tr>
<td>FWS_137_POL</td>
<td>35.70</td>
<td>4.0</td>
</tr>
</tbody>
</table>

*Table 2.4*: Details the DNA quantities obtained from each individual and the volume of Tagmentation enzyme used for each individual library preparation. Any reactions where less than the recommended 5 µl of Tagmentation enzyme were used, the volume was made up to 5 µl with EB buffer so the final reaction volume was not affected.
Figure 2.5: Examples of successful (top) and poor (bottom) library fragment size distributions. The top figure is an example of a successful human genomic DNA library size distribution (taken from the Illumina Nextera manual). The bottom figure is an example of one of the poorest library size distributions from this study that was not used.

2.3.4 Bioinformatic pipeline overview

The bioinformatic pipeline used here was developed to create datasets consisting of short sequence blocks aligned across a small number of individuals per species. The main stages involved are outlined in Figure 2.6, and I summarise each step briefly here. Each step is considered in more detail in sections 2.3.4.1-2.3.4.5.

Quality control of the raw reads is required to remove any poor quality bases from each read and to ensure that all the adapter sequences used during the sequencing process have been removed. If the libraries include fragments that are shorter than the paired read length (i.e. here ~250-300 bp) some of the paired end reads will overlap. In these cases it is beneficial to merge these reads, creating longer reads that will assist in the downstream assembly and mapping stages.

When using whole organisms to generate genome sequences there is a risk of not just generating sequences from the target organism. For example, Wolbachia bacteria are known to be common endosymbionts in fig wasps (Haine and Cook, 2005) and a diverse range of fungal species have been identified from a genome assembly of the fig wasp species Ceratosolen solemsi (Niu et al., 2015). There is also a risk of
The presence of contaminant reads would be detrimental to the downstream goal of assessing the evolutionary relationships of the targeted fig wasps. To assess the level and type of contamination present in each sample, I ran them through the blobology pipeline (Kumar et al., 2013) and used these results to filter out potential contaminant reads from the data. The blobology pipeline creates Taxon-Annotated-GC-Coverage plots (Blobplots) that use BLAST searches to identify aligned reads at a specified taxonomic level (here phylum). The blobs on each plot represent each contig (the size of each blob is proportional to the length of each contig), which is coloured according to its phylum designation and plotted using its GC content (x-axis) and coverage (y-axis). Contigs from different phyla (and particularly eukaryotes versus prokaryotes) differ in their GC:AT content ratio, and so form clusters of points (blobs) on the plot. Sequences identified as originating in Wolbachia or other bacterial symbionts can thus be visualised, and the contamination quality of the sample assessed. Contaminant sequences can then be excluded. Example blobplots can be seen in Figure 2 of Kumar et al (2013).

A reference assembly is required to be able to identify variable sites (i.e. call variants) from any genomic dataset, which can be difficult when working with non-model organisms. To address this I used the reads from all the individuals of the same species to build a de novo reference assembly. The per individual reads were then aligned to the appropriate assembly and variants called using the GATK variant calling pipeline (McKenna et al., 2010; Depristo et al., 2011; Van Der Auwera et al., 2013). The pipeline uses a range of mapping, base and coverage quality metrics to establish the probability of a base in an aligned read being genuinely different to the reference base.

Repetitive DNA is present in the genomes of organisms from across the kingdoms of life (Treangen and Salzberg, 2011) and the percentage of the genome covered is known to vary in insects (Wang et al., 2008). It is important to attempt to identify repetitive regions in target organisms as their presence can cause assembly and mapping errors especially when using short read lengths (Treangen and Salzberg, 2011). For this study it is also important to mask these regions, as repetitive DNA elements do not conform to the mutation model assumed in the downstream modelling procedures.

The final stage in the pipeline involves filtering of variant calls to obtain only those that are present in all the individuals of a species, resulting in a set of aligned sequence blocks suitable for population genomic analysis.
Figure 2.6: Flowchart detailing the main stages of the bioinformatic processing pipeline.

2.3.4.1 Quality control and processing of sequence reads
The initial quality of the sequence reads was assessed using Fast-QC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were quality trimmed at a base quality score of 20 (sliding window 15:20 and min length 50) and adapter trimmed using Trimomatic version 0.32 (Bolger et al., 2014). The trimmed reads were again run through Fast-QC, which highlighted that some adapter sequences remained. To remove these sequences the reads were run through the program Cutadapt (Martin, 2011). As the reads came from three lanes this part of the processing was carried out at different times. As a result the F. watkinsiana samples and the first HiSeq lane were processed using Cutadapt version 1.4.2 and the second lane was processed using Cutadapt version 1.8.3. The earlier version required any reads of length zero (i.e. reads that had been trimmed to leave no sequence remaining with their read names still present in the file) to be removed
prior to the next processing step. This was done using two custom shell scripts written by J. Hearn. Paired-end reads were merged using PEAR version 0.9.0 (Zhang et al., 2014). Fast-QC was re-run on the filtered reads to check the success of the filtering process.

### 2.3.4.2 Blobplots

To produce the blobplots, I created de novo meta-assemblies using the short read de Bruijn graph assembler Velvet (version 1.2.10) and a k-mer length of 31 (Zerbino and Birney, 2008). Combining the reads from all individuals of the same species and assembling them together created the meta-assemblies. This approach increases the overall coverage and provides the best possible reference assembly for each species (see Table 2.5 for details on the numbers of individuals contributing to each meta-assembly). Statistics to assess these assemblies were generated using a perl script (J. Hearn) and by running them through the Core Eukaryotic Genes Mapping Approach (CEGMA) program (version 2.5) (Parra et al., 2007; Parra et al., 2009). CEGMA is a method for assessing the completeness of a genome assembly using a well-characterised set of conserved eukaryotic genes. These assembly statistics are shown in Table 2.5. The filtered reads were mapped to the meta-assemblies using Bowtie2 version 2.2.3 (Langmead and Salzberg, 2012). Four BLAST approaches were used to assess the meta-assemblies for taxonomic matches. The fast protein aligner Diamond (version 0.7.9) (Buchfink et al., 2015) was used alongside three BLAST (version 2.2.29) searches: 1) against the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/), 2) against the draft genome of the Agaonidae pollinating fig wasp, Ceratosolen solmsi (Xiao et al., 2013), and 3) against the draft genome of the pteromalid parasitoid wasp Nasonia vitripennis (Werren et al., 2010). The two latter searches were done to utilise the genomes of the most closely related individuals to the species used in this study. The outputs from Diamond (.daa files) were converted to blobplot compatible files using a perl script written by D. Laetsch. Blobplots were produced using the python scripts deposited at https://github.com/DRL/blobtools-light, written by D. Laetsch.
Table 2.5: Details of the number of individuals that were used in each assembly and a set of statistics describing the Velvet assemblies. The assemblies are labelled according to the tree of origin (e.g. FR for Ficus rubiginosa) and the genus of wasp (e.g. SYC for Syoscapter, in the case of Pleistodontes, POL). The contig N50 measure of assembly quality states that 50% of bases in the assembly are contained in contigs greater than or equal to this value. The CEGMA ‘Complete’ % completeness indicates the percentage of the 248 CEGs (core eukaryotic genes) where >70% of the protein length was covered by the genome assembly. The CEGMA ‘Partial’ % completeness indicates the percentage of the 248 CEGs where <70% of the protein length was covered by the genome assembly (inclusive of the complete matches).
2.3.4.3 Contaminant read removal

Most of the blobplots show most contigs to be assigned to Arthropoda, or to be unassigned to any single phylum (Appendix C). They also show that all assemblies only contain a low level of contamination, i.e. contigs with BLAST hits to other phyla. Following these results I decided to choose an arbitrary cut-off value and only remove reads mapping to contigs if the assembly had >1200 contigs showing a best BLAST hit to a non-arthropod phylum. Using the blobplot output file as input, I wrote a custom shell script to remove these reads from both the trimmed and the merged read files. The resultant read files were used in all further processing. Proteobacteria matching reads were removed from the F. rubiginosa-associated species Pleistodontes imperialis, Sycoscapter long, Philotrypesis, Herodotia and Eukobelea and Proteobacteria and Ascomycota matching reads were removed from the F. watkinsiana-associated species Pleistodontes nigriventris.

2.3.4.4 Assembly, mapping, variant calling and repeat masking

2.3.4.4.1 Assembly and quality assessment

Unlike the quick Velvet assemblies made for use in the blobplots, the final assemblies were made using the program SPAdes (version 3.6.2) (Bankevich et al., 2012). SPAdes is another de Bruijn graph assembler but, unlike Velvet, it uses a multi-sized approach meaning it can use differing k-mer lengths along the genome. The choice of k-mer length will have a large effect upon the resulting assembly especially when it contains regions of variable coverage (Bankevich et al., 2012). This use of varying k-mer lengths and the SPAdes error correction tool improves assembly quality compared to other assemblers, e.g. Velvet, especially when using low coverage data (Bankevich et al., 2012). The filtered, merged read files of all individuals from the same species were combined prior to assembly for the reasons stated above. The quality of these meta-assemblies was assessed using a perl script and CEGMA as above and also the program BUSCO (Benchmarking Universal Single-Copy Orthologs) version 1.1b1 (Simao et al., 2015) (Tables 2.6 and 2.7). BUSCO is a method (similar to CEGMA) for assessing genome completeness in terms of gene content. It uses a set of single copy orthologs compiled from hundreds of genomes from across the tree of life, with BUSCO sets defined for six major phylogenetic clades (Simao et al., 2015). The assemblies were run through BUSCO using the Arthropoda BUSCO set (http://busco.ezlab.org/). From here-on in I will refer to the SPAdes assemblies as reference assemblies.
Table 2.6: Details the number of individuals that were used in each assembly and a set of statistics describing the *SPAdes* assemblies. The assemblies are labelled according to the tree of origin (e.g. FR for *Ficus rubiginosa*) and the genus of wasp (e.g. SYC for *Sycoscapter*, in the case of *Pleistodontes*, POL). The contig/scaffold N50 measure of assembly quality states that 50% of bases in the assembly are contained in contigs/scaffolds greater than or equal to this value.

<table>
<thead>
<tr>
<th></th>
<th>SPAdes assemblies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of individuals</td>
<td>No. of contigs (&gt;=200bp)</td>
</tr>
<tr>
<td>FO_POL</td>
<td>4</td>
<td>65,958</td>
</tr>
<tr>
<td>FO_SYC</td>
<td>4</td>
<td>204,520</td>
</tr>
<tr>
<td>FO_PHIL</td>
<td>4</td>
<td>152,168</td>
</tr>
<tr>
<td>FR_POL</td>
<td>6</td>
<td>85,907</td>
</tr>
<tr>
<td>FR_SYC_long</td>
<td>6</td>
<td>54,476</td>
</tr>
<tr>
<td>FR_SYC_short</td>
<td>4</td>
<td>69,048</td>
</tr>
<tr>
<td>FR_PHIL</td>
<td>8</td>
<td>202,642</td>
</tr>
<tr>
<td>FR_WAT</td>
<td>6</td>
<td>60,725</td>
</tr>
<tr>
<td>FR_HERO_2nd</td>
<td>7</td>
<td>136,987</td>
</tr>
<tr>
<td>FR_EUKO</td>
<td>4</td>
<td>90,076</td>
</tr>
<tr>
<td>FW_POL</td>
<td>4</td>
<td>139,731</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>NO. of individuals</th>
<th>GC content</th>
<th>CEGMA - complete % completeness</th>
<th>CEGMA - partial % completeness</th>
<th>BUSCO Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>FO_POL</td>
<td>4</td>
<td>0.282</td>
<td>81.85</td>
<td>96.37</td>
</tr>
<tr>
<td>FO_SYC</td>
<td>4</td>
<td>0.384</td>
<td>85.48</td>
<td>97.18</td>
</tr>
<tr>
<td>FO_PHIL</td>
<td>4</td>
<td>0.396</td>
<td>91.94</td>
<td>98.79</td>
</tr>
<tr>
<td>FR_POL</td>
<td>6</td>
<td>0.294</td>
<td>86.29</td>
<td>96.37</td>
</tr>
<tr>
<td>FR_SYC_long</td>
<td>6</td>
<td>0.384</td>
<td>94.76</td>
<td>98.79</td>
</tr>
<tr>
<td>FR_SYC_short</td>
<td>4</td>
<td>0.386</td>
<td>93.95</td>
<td>98.39</td>
</tr>
<tr>
<td>FR_PHIL</td>
<td>8</td>
<td>0.397</td>
<td>94.35</td>
<td>98.39</td>
</tr>
<tr>
<td>FR_WAT</td>
<td>6</td>
<td>0.386</td>
<td>95.56</td>
<td>98.39</td>
</tr>
<tr>
<td>FR_HERO_2nd</td>
<td>7</td>
<td>0.369</td>
<td>70.97</td>
<td>93.55</td>
</tr>
<tr>
<td>FR_EUKO</td>
<td>4</td>
<td>0.438</td>
<td>90.32</td>
<td>98.79</td>
</tr>
<tr>
<td>FW_POL</td>
<td>4</td>
<td>0.296</td>
<td>93.15</td>
<td>99.19</td>
</tr>
</tbody>
</table>

Table 2.7: Details the number of individuals that were used in each assembly, the GC content and the CEGMA and BUSCO scores of each SPAdes assemblies. The assemblies are labelled according to the tree of origin (e.g. FR for Ficus rubiginosa) and the genus of wasp (e.g. SYC for Syoscepter, in the case of Pleistodontes, POL). The CEGMA ‘Complete’ % completeness indicates the percentage of the 248 CEGs (core eukaryotic genes) where >70% of the protein length was covered by the genome assembly. The CEGMA ‘Partial’ % completeness indicates the percentage of the 248 CEGs where <70% of the protein length was covered by the genome assembly (inclusive of the complete matches). The BUSCO scores represent the percentages of complete (C), duplicated (D), fragmented (F) and missing (M) BUSCOs (Benchmarking sets of Universal Single-Copy Orthologs) found in the assembly. ‘n’ represents the number of BUSCOs searched (the Arthropoda set).
2.3.4.4.2 Mapping and variant calling pipeline

The filtered, paired reads from each individual were mapped back to their respective reference assemblies using the Burrows-Wheeler Aligner (BWA) version 0.7.10 (Li and Durbin, 2009). BWA is the recommended aligner for use with the Genome Analysis Tool Kit (GATK) variant calling pipeline (McKenna et al., 2010; Depristo et al., 2011; Van Der Auwera et al., 2013). To prepare the reference assemblies for use in the GATK pipeline, FASTA file indexes and sequence dictionaries were created using samtools (Li et al., 2009) faidx (version 1.2) and the picard (https://broadinstitute.github.io/picard/index.html) tool CreateSequenceDictionary (version 1.141) respectively. The BAM (mapping) files were sorted and merged to create a single BAM file per species using the picard tool MergeSamFiles. PCR duplicates were removed from these merged files using the picard tool MarkDuplicates. During initial runs of the GATK HaplotypeCaller tool I discovered that it was failing to run on the smallest contigs when all the data were included. For this reason I decided to limit the analysis to contigs greater than 500 bp in length. To do this the GATK tools require an interval list containing all the contig IDs to be included in the analysis. I created the interval lists using a custom shell script. Reads were realigned around indels using GATKs IndelRealigner tools (version 3.5.0) and the manufacturers instructions. Base quality score recalibration (BQSR) was performed using the bootstrap technique recommended by GATK developers as there is no known SNP database available for these species. Variants were called using GATKs HaplotypeCaller and were hard filtered (i.e. using a set of filtering cut-off values) as recommended in their best practice guidelines for datasets where the variant quality score recalibration (VQSR) is not possible. As the sequenced males are haploid the ploidy was set to 1 and the ‘-emit variants only’ option was used when running the HaplotypeCaller. Finally the SNP and indel VCF (variant call format) files were merged using the picard tool MergeVCF. The pipeline results in three files per species: SNP only, indel only, and a combined VCF file.

2.3.4.4.3 Masking repetitive regions

As there are no repeat databases available for these target species I used RepeatScout (version 1.0.5) (Price et al., 2005) to create a library of repetitive regions identified from each of the de novo reference assemblies. RepeatMasker (version open-4.0.6) (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>) was then used to mask these regions across the assemblies. The eleven RepeatScout libraries obtained from each of the species assemblies were combined along with the recommended RepeatMasker library. This
combined library of repeats was used as the input for the masking process in RepeatMasker. RepeatMasker outputs an assembly file with repetitive elements represented as runs of Ns and an annotation file listing the positions of the repeats. I used this annotation file to create a BED file of contig IDs and repeat positions for use in downstream processing steps. The script uses the bedtools tool (Quinlan and Hall, 2010) mergeBed (version 2.23.0) to create the BED file. All assemblies were run in RepeatMasker using default parameters with the exception of the assembly for P. nigriventris (ex F. watkinsiana), which was split into 10 FASTA files and run in RepeatMasker quick mode due to the program not completing successfully when running the whole assembly.

2.3.4.5 VCF filtering
The GATK tool CallableLoci assesses whether sites are callable based upon the coverage and mapping quality at each locus. To class a base as callable default values for parameters were used except for the following: minimum base quality of 10 (1 in 10 probability that the base was called wrong), minimum mapping quality of 20 (1 in 100 probability that the alignment is wrong), minimum read depth of 2 (at least two reads covering each loci). Each base quality score recalibrated (BQSR) BAM file was subsampled to pull out per individual BAM files. These were run through CallableLoci to output a BED file listing the callable regions within each alignment. The BQSR BAM files were also run through QualiMap 2 (García-Alcalde et al., 2012; Okonechnikov et al., 2016), a program that assesses alignment quality and outputs summaries of the main properties of the data (e.g. ACGT content, coverage and mapping quality metrics). Bedtools subtractBed was used to remove the repeat regions (using the RepeatMasker BED files created above) from the callable loci BED files. Overlapping positions present across all individuals of the same species were extracted using bedtools multiIntersectBed. Only sites deemed ‘Callable’ or ‘Low Coverage’ by CallableLoci were included in this final filtered BED file. This BED file was used to filter the VCF files (generated from the GATK pipeline) for callable variable sites using bcftools (Li, 2011) version 1.2.

2.3.5 Genome-wide concatenated phylogenies
The genome-wide SNP datasets for each individual were used to summarise the average distance between the individuals of each species. I generated an aligned variant-only FASTA file per species using a series of custom python and bash scripts. The scripts use the filtered per species SNP-only VCF files as input to call each per individual variant site based on their genotypes. The FASTA files were
converted to NEXUS format using the `seqmagick` utility (http://seqmagick.readthedocs.io/en/latest/). The NEXUS files were converted to PHYLIP format using the sequence converter at http://sequenceconversion.bugaco.com/converter/biology/sequences/. The alignments were used to produce maximum likelihood (ML) trees in `PhyML 3.0` (Guindon et al., 2010). The HKY85 model of substitution was used for each dataset. Support values were estimated using the aBayes method (Anisimova et al., 2011) implemented in `PhyML 3.0`. The unrooted trees are displayed in the radial layout as no outgroups were available for the SNP datasets in this form. The trees were edited and exported using `FigTree` version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) and further edited using `Inkscape` 0.91 (https://inkscape.org/en/).
2.4 Results

2.4.1 Selection of individuals for HiSeq sequencing

DNA was extracted from a total of 157 individuals, all of which were DNA barcoded with at least one set of primers (Table 2.2). Individuals of each barcoded species were built into a cytb phylogeny, as described in the Methods. In the sections below I present DNA barcoding results, and the resulting selection of samples for library generation, by species within each sampled wasp genus.

**Pleistodontes**: Darwell (2013) described five putative *P. imperialis* (*ex* *F. rubiginosa*) species based on mitochondrial cytb and CO1 and nuclear ITS2 loci. The cytb sequences from several individuals (those ending Sp[1-5]) from that study were used to populate the *Pleistodontes* cytb phylogeny (Figure 2.7) in this study. The concordance between the clades and the distributional information presented in Darwell (2013) and the current study was used to confirm the identity of the *P. imperialis* individuals sequenced here. The target *P. imperialis* (sp. 1) forms a strongly supported clade (Figure 2.7, clade 1) within which two clades (one poorly supported) suggest differentiation between individuals sampled from the North and South. All samples were chosen arbitrarily from these two clades with the first two Sydney samples failing to produce decent library traces forcing a second pair to be attempted. The six *P. imperialis* sp. 1 (*ex* *F. rubiginosa*) individuals sequenced were FRN_J7_POL, FRN_364_POL, FRS_234_POL, FRS_511_POL, FRSYD_1065_POL and FRSYD_2884_POL. There are two described pollinator species associated with *F. obliqua*, *P. xanthocephalus* and *P. greenwoodi* (Lopez-Vaamonde et al., 2002, Chapter 1). The extra individuals used to populate the cytb phylogeny (Figure 2.7) were identified as *P. xanthocephalus* (T. Sutton, personal communication). The samples sequenced for this study form a strongly supported clade (clade 6) with these previously identified *P. xanthocephalus* individuals leading to the conclusion they too were *P. xanthocephalus* despite the southern sampling. Libraries were made for individuals FON_255_POL and FON_409_POL but both were extremely poor resulting in the following four individuals being sequenced: FON_403_POL, FON_398_POL, FOS_183_POL1 and FOS_187_POL. The four *P. nigriventris* (*ex* *F. watkinsiana*) individuals in the cytb phylogeny form a clade (clade 7) with strong support (Figure 2.7). As there are currently no known cryptic pollinator species associated with *F. watkinsiana* figs (Dixon, 2003) I was confident in the morphological (carried out by J. Cook) and molecular identification of these individuals. The library preparations for these four *P. nigriventris* individuals were successful and so they were sequenced on the HiSeq.
Figure 2.7: Maximum likelihood phylogeny reconstructed from mtDNA cytb regions for *Pleistodontes* individuals from *F. rubiginosa*, *F. obliqua* and *F. watkinsiana*. aBayes node support values are indicated. Samples from this study are indicated in red and those chosen for high-throughput sequencing are indicated by the stars (*).

*Sycoscapter*: Segar (2011) described three putative *Sycoscapter* species associated with *F. rubiginosa*, a result reiterated by Darwell (2013). The first study was based on
sequences from mitochondrial cytb and nuclear 28S and ef1a loci whereas the second used cytb and nuclear ITS2 sequences. Both showed a morphological distinction using ovipositor length in females (long and short) and showed that the short species could be further divided into two groups, sampling locations suggesting they have different distributions. The *Sycoscapter* (*ex F. rubiginosa*) individuals form a strongly supported clade (clade [1,2]) in the cytb phylogeny reconstructed for the current study (Figure 2.8). This clade is subsequently divided into two strongly supported clades (clades 1 and 2) which, based upon the previous designations of the additional sequences that were used to populate the tree, correspond to the long and short forms respectively. Six individuals were chosen from the *Sycoscapter* long clade (clade 1), two from each population, each of the pairs from different sampling sites. The six individuals sequenced were FRN_274_SYC1, FRN_350_SYC, FRS_32_SYC, FRS_397_SYC, FRSYD_89_SYC and FRSYD_2873_SYC. Four individuals were chosen from the *Sycoscapter* short clade (clade 2), which has nested within it two clades seemingly representing the two sister species, sp.1 and sp.2, based upon previous knowledge of their distributions. The four *Sycoscapter* short individuals sequenced were FRN_332_SYC, FRS_238_SYC, FRS_508_SYC and FRSYD_1941_SYC. Segar (2011) report two closely related *Sycoscapter* species associated with *F. obliqua* based on molecular analyses (markers as stated above) but that show no obvious subgroups based on ovipositor lengths in females. The study reports considerable morphological and molecular variation across the sampled individuals suggesting no clear split in distribution of the two proposed species. The molecular analyses suggesting two species were supported by the results of Darwell (2013). The *Sycoscapter* (*ex F. obliqua*) individuals form a poorly supported clade in the current cytb phylogeny (Figure 2.8, clade 3). The strongly supported clade within clade 3, groups the individuals together with the sp.1 designated individual from the Darwell (2013) study and indicates a split between North and South individuals. Due to library preparation failures, the individual FON_TrinBeach_SYC was sequenced despite its poorly supported location within clade 3. The three other individuals sequenced were FON_339_SYC, FOS_186_SYC and FOS_199_SYC.
Figure 2.8: Maximum likelihood phylogeny reconstructed from mtDNA cytb regions for *Sycoscapter* individuals from *F. rubiginosa* and *F. obliqua*. aBayes node support values are indicated. Samples from this study are indicated in red and those chosen for high-throughput sequencing are indicated by the stars (*).

**Philotrypesis**: Philotrypesis females can be distinguished by body colour (black and yellow) and were previously presented as two *F. rubiginosa*-associated species (Segar, 2011). Darwell (2013) however further characterised each colour form into two groups using cytb and ITS2 sequences, proposing there are four *F. rubiginosa*-associated Philotrypesis species (yellow 1 and 2, black 1 and 2). The Philotrypesis (ex *F. rubiginosa*) individuals are distributed across three strongly supported clades (Figure 2.9, clades 1, 3 and 4) and one with moderate support (clade 2). Under the assumption the cytb sequences (Darwell, 2013) used to populate the phylogeny
(Figure 2.9) are appropriate to identify the species in each clade, the North population individuals are *Philotrypesis* yellow sp.2 (clade 2), the South individuals are *Philotrypesis* yellow sp.1 (clade 1) or *Philotrypesis* black sp.1 (clade 4) and the Sydney individuals are *Philotrypesis* black sp.1 (clade 4) or black sp.2 (clade 3). In the end eight individuals from across the tree were sequenced: FRN_310_PHIL, FRN_407_PHIL, FRS_40_PHIL, FRS_247_PHIL, FRS_110_PHIL, FRS_511_PHIL, FRSYD_165_PHIL and FRSYD_1822_PHIL. The decision to sequence eight instead of six individuals was made in order to have two representatives from each clade however due to barcoding and library preparation difficulties, three individuals were sequenced from clade 4 and only one from clade 3. Wasps of the *Philotrypesis* (*ex F. obliqua*) genus are also found in two morphologically distinct forms (yellow and black) (Segar, 2011) however Darwell (2013) proposed the yellow form was composed of two distinct groups. This distinction of the yellow species was based only upon ITS2 sequences of which there were only two samples. The results of the cytb analysis grouped one of these samples with the black individuals (there was no cytb sequence for the other sample). Therefore, the sequences (Darwell, 2013) used to populate the cytb phylogeny (Figure 2.9) represent the black (clade 6) and yellow (clade 5) species as previously defined (Segar, 2011). Both the *Philotrypesis* (*ex F. obliqua*) clades (5 and 6) are strongly supported and suggest a North/South split between the two. Two individuals from each clade were sequenced: FON_340_PHIL, FON_PDoug_PHIL, FOS_186_PHIL and FOS_188_PHIL1.
Figure 2.9: Maximum likelihood phylogeny reconstructed from mtDNA cytb regions for Philotrypesis individuals from *F. rubiginosa* and *F. obliqua*. Bayes node support values are indicated. Samples from this study are indicated in red and those chosen for high-throughput sequencing are indicated by the stars (*).

**Watshamiella**: The wasps associated with *F. rubiginosa* from this genus are found as two genetically distinct forms, the females of which are morphologically distinguishable based on ovipositor length (long and short) (Segar, 2011; Darwell, 2013). One cytb sequence from each (Darwell, 2013) were used to identify each species in the cytb phylogeny (Figure 2.10). The strongly supported clade 1 seemingly represents *Watshamiella* long individuals. The sequences from the current study in this clade show they all come from Sydney. Clade 2 (*Watshamiella* short) is poorly supported but contains representatives from each population. Six individuals were sequenced from clade 2: FRN_299_WAT, FRN_337_WAT, FRS_239_WAT, FRS_514_WAT, FRSYD_567_WAT and FRSYD_1905_WAT1.
Figure 2.10: Maximum likelihood phylogeny reconstructed from mtDNA cytb regions for Watshamiella individuals from F. rubiginosa. Bayes node support values are indicated. Samples from this study are indicated in red and those chosen for high-throughput sequencing are indicated by the stars (*).

Herodotia: Segar (2011) reported a single species of wasp from the genus Herodotia associated with F. rubiginosa however Darwell (2013) suggested there were three genetically distinct species based upon analysis of cytb and ITS2 loci. A cytb sequence from each of these proposed species (Darwell, 2013) was used to populate the current cytb phylogeny (Figure 2.11). Clade 1 is strongly supported and contains one North and one South individual but also contains sequences derived from Herodotia (ex F. obliqua) individuals (Darwell, 2013). Clades 2 and 3 are poorly supported and contain individuals from all three populations. The poor support for clades 2 and 3 essentially collapses them resulting in the smaller clades nested
within them becoming several strongly supported clades that have no clear species designation. Due to this uncertainty and the expectation these large wasps might yield higher quality sequence data, nine individuals, three from each population, were chosen from across the tree to sequence: FRN_261_HERO, FRN_372_HERO, FRN_349_HERO, FRS_110_HERO, FRS_239_HERO, FRS_517_HERO, FRSYD_381_HERO, FRSYD_293_HERO and FRSYD_1964_HERO. Individual FRS_239_HERO failed to produce any sequence data and so was replaced with individual FRS_244_HERO on the second lane.

Figure 2.11: Maximum likelihood phylogeny reconstructed from mtDNA cytb regions for *Herodotia* individuals from *F. rubiginosa*. aBayes node support values are indicated. Samples from this study are indicated in red and those chosen for high-throughput sequencing are indicated by the stars (*).
**Eukobelea**: Wasps of the genus *Eukobelea* associated with *F. rubiginosa* have previously been described as a single species (Segar, 2011) however Darwell (2013) reported two genetically distinct but morphologically cryptic species (1 and 2) with predominantly North/South distributions. Two cytb sequences from each species (1 and 2) (Darwell, 2013) were used to populate the cytb phylogeny (Figure 2.12). Clade 1 is strongly supported and groups the two South individuals together with the sp.1 samples. However clade 2 has moderate support and only contains two of the three North individuals along with the sp.2 samples. The third North individual does not group with any of the other sequences. The two North individuals in clade 2 are from the same site leading to the following four being chosen for sequencing: FRN_406_EUKO, FRN_310_EUKO, FRS_41_EUKO and FRS_239_EUKO.

**Figure 2.12**: Maximum likelihood phylogeny reconstructed from mtDNA cytb regions for *Eukobelea* individuals from *F. rubiginosa*. aBayes node support values are indicated. Samples from this study are indicated in red and those chosen for high-throughput sequencing are indicated by the stars (*).
From the barcoded individuals, 78 libraries were made, 60 of which were sequenced on the Illumina HiSeq. Table 2.1 details the locations and designated populations of all those sequenced.

### 2.4.2 DNA quantification and library preparation

After Qubit-ing, 70% of the samples did not contain 50 ng DNA in the elution volume. This was not surprising given the small size of the wasps but was a potential worry as the library protocol recommended 50 ng as the starting concentration and so whether it would be successful with less was unclear. Table 2.4 details the quantity of DNA obtained for each sample. Unsurprisingly in general the smaller the wasp the smaller the DNA quantity obtained but there was variation in the quantities obtained within each species. Despite the low quantities of DNA, I generated libraries using the small concentrations available. There was a lot of variation in the fragment size distributions across libraries. The majority were either flat ‘humped’ (as in Figure 2.13) or showed double peaks around the smallest and largest fragment sizes (as in Figure 2.13). I found no clear pattern in the distributions seen across species, or any clear relationship between DNA quantity and distribution shape. Despite these less than ideal library distributions, only one sample out of 60 failed to yield any data altogether.

![Example of ‘humped’ and double peaked bioanalyser traces](image)

**Figure 2.13:** Examples of ‘humped’ (top) and double peaked (bottom) bioanalyser traces.
2.4.3 Assessing contamination and assembly/mapping quality

Visual inspection of the blobplots suggests that the pollinating fig wasps (*Pleistodontes ex F. watkinsiana*, *F. obliqua* and *F. rubiginosa*) assembled more successfully than any of the NPFWs (for the full range of plots see Appendix C). Overall, across all three species, pollinating fig wasps have fewer contigs (blobs), generally higher coverage and more contigs at around 20-30% GC content. This last observation that pollinators are particularly AT rich is consistent with the low GC content revealed from the only sequenced genome of a pollinating fig wasp (Xiao et al., 2013). All of the NPFWs show larger diffuse ‘clouds’ of small blobs that vary more in coverage and GC content. In all of the species, except the *Eukobelea (ex F. rubiginosa)* wasps, the largest proportion of contigs are designated Arthropoda. The three wasp species (*Pleistodontes*, *Sycoscapter* and *Philotrypesis*) from *F. obliqua* show very little, if any, contamination. The majority of contigs were labelled as Arthropoda or ‘no hit’ with only very small numbers being labelled as other phyla. The same pattern was seen in the *Sycoscapter* short and *Watshamiella* individuals (ex *F. rubiginosa*). In the other *F. rubiginosa* species (*Pleistodontes imperialis*, *Sycoscapter* long, *Philotrypesis*, *Herodotia* and *Eukobelea*) however, a much higher number of contigs were labelled as Proteobacteria (>1200 contigs). The plots show which individual has contributed these contaminant reads to the reference assembly. For example, each *P. imperialis* individual shows a cluster of Proteobacteria-designated contigs but individual FRN_J7_POL stands out as it shows a large cluster of high coverage Proteobacteria-designated contigs (Figure 2.14). The *P. nigriventris (ex F. watkinsiana)* assembly was the only one to show a high number (>1200 contigs) of contigs labelled Ascomycota as well as a high number labelled as Proteobacteria (Figure 2.15). The two northern individuals contribute most of the Proteobacteria reads to the assembly with the southern individual FRS_03-250_POL contributing most of the Ascomycota reads. A general pattern amongst all the species that show Proteobacteria contamination is that the northern individuals tend to be the ones contributing the highest amount.
Population genomics of a fig wasp community

FRN_J7_POL
Figure 2.14: Blobplots for each of the six *P. imperialis* (*ex F. rubiginosa*) individuals. The blobs represent contigs, the size of each corresponding to contig length, and are coloured according to phylum designation. The blue blobs in these plots represent Proteobacteria-designated contigs. The numbers in each key correspond to: the total number of contigs designated to that phylum, the total length of the contigs (span) designated to that phylum, the average length of the contigs designated to that phylum. The phyla are ordered by the total contig length, from highest to lowest.
Population genomics of a fig wasp community

FWN_01-8_POL
Population genomics of a fig wasp community

FWS_03-250_POL
Population genomics of a fig wasp community

Figure 2.15: Blobplots for each of the four *P. nigriventris* (*ex F. watkinsiana*) individuals. The blobs represent contigs, the size of each corresponding to contig length, and are coloured according to phylum designation. The blue blobs in these plots represent Ascomycota-designated contigs and the pink blobs represent Proteobacteria-designated contigs. The numbers in each key correspond to: the total number of contigs designated to that phylum, the total length of the contigs (span) designated to that phylum, the average length of the contigs designated to that phylum. The phyla are ordered by the total contig length, from highest to lowest.
Two of the blobplots representing the *Herodotia* (*ex* *F. rubiginosa*) individuals, FRN_261_HERO and FRS_110_HERO (Figure 2.16), looked particularly poor in the spread of the large numbers of very small contigs. This suggested these two individuals are sufficiently different from the others (in agreement with the cytb phylogeny) to result in a poor assembly and subsequent mapping. I decided to remove these two individuals from the assembly, reassemble and remake the blobplots for this species (Appendix C). This improved the blobplots for the remaining individuals and improved the assembly statistics for the underlying reference assembly (see Table 2.5 FR HERO/FR HERO 2nd).
Figure 2.16: Blobplots for two of the Herodotia (ex F. rubiginosa) individuals. The blobs represent contigs, the size of each corresponding to contig length, and are coloured according to phylum designation. The pink blobs in these plots represent Proteobacteria-designated contigs. The numbers in each key correspond to: the total number of contigs designated to that phylum, the total length of the contigs (span) designated to that phylum, the average length of the contigs designated to that phylum. The phyla are ordered by the total contig length, from highest to lowest.

Following removal of the putative contaminant reads, each species was re-assembled using SPAdes. Basic assembly statistics (e.g. N50, mean contig length, max contig length and CEGMA scores) summarising both the Velvet and SPAdes
assemblies are shown in Tables 2.5, 2.6 and 2.7. For all assembly quality measures the SPAdes assemblies are an improvement on the equivalent Velvet ones however the SPAdes assembly quality varies across each species sampled. Using the scaffold N50 (50% of bases in the assembly are contained in scaffolds greater than or equal to this value) as a measure of quality, they range from 2,966 bp to 35,331 bp. However, when the highest scoring assembly is ignored the range is not as large with the second highest N50 value being 12,182 bp. The three pollinating species (Pleistodontes ex F. watkinsiana, F. obliqua and F. rubiginosa) have lower GC contents compared to the NPFWs (~29% compared to ~39%). The only other pollinating fig wasp and parasitoid wasp genomes to have been sequenced (Ceratosolen solmsi and Nasonia vitripennis) have GC contents of 30.4% (Xiao et al., 2013) and 40.6% (Werren et al., 2010), respectively. These species level meta-assemblies were used as reference assemblies when aligning each individual’s paired reads. Table 2.8 shows some basic mapping statistics for each individual alignment. There is a lot of variation in the total number of reads per individual but the percentage of reads aligned to the reference is always above 99%. The number of reads that map to their particular reference with a mapping quality score of 20 or higher (i.e. a 1 in 100 chance of misalignment) as a percentage of the total number of reads aligned is above 90% for all individuals except those of the Philotrypesis (ex F. rubiginosa) wasps. These range from 88-90% reads aligned with a quality score of 20 or above. Mean coverage per meta-assembly is shown in Table 2.9 alongside a per individual average. The per individual averages fall short of the 6-fold coverage target when preparing the library pools, which will at least in part be due to not receiving the expected number of reads off each sequencing lane. The pooling volumes were calculated under the assumption each lane would yield ~400 million 125 bp reads (50 Gb) however the number of reads received was less than this for both lanes and especially so for the first lane (~345 million and ~398 million reads from the first and second lanes respectively). The percentage of each assembly that was masked during the repeat masking process ranged from 8-27% (Table 2.9). The number of SNPs called for each species through the GATK variant calling pipeline (Table 2.9) covers a large range, from 312,329 in Sycoscapter long (ex F. rubiginosa) to 3,016,975 in Eukobelea (ex F. rubiginosa).
<table>
<thead>
<tr>
<th>Individual ID</th>
<th>Total number of reads (including both of the pair)</th>
<th>Percentage of reads aligned (including both of the pair)</th>
<th>Reads that mapped with Q20 or higher as percentage of all aligned reads*</th>
</tr>
</thead>
<tbody>
<tr>
<td>FON_398_POL</td>
<td>12,749,732</td>
<td>99.89</td>
<td>98.00</td>
</tr>
<tr>
<td>FON_403_POL</td>
<td>13,241,975</td>
<td>99.86</td>
<td>97.93</td>
</tr>
<tr>
<td>FOS_183_POL1</td>
<td>13,322,377</td>
<td>99.84</td>
<td>97.89</td>
</tr>
<tr>
<td>FOS_187_POL</td>
<td>14,297,657</td>
<td>99.83</td>
<td>98.17</td>
</tr>
<tr>
<td>FON_339_SYC</td>
<td>10,729,830</td>
<td>99.44</td>
<td>92.96</td>
</tr>
<tr>
<td>FON_TrinBeach_SYC</td>
<td>13,599,521</td>
<td>99.77</td>
<td>93.92</td>
</tr>
<tr>
<td>FOS_186_SYC</td>
<td>10,253,814</td>
<td>99.85</td>
<td>93.10</td>
</tr>
<tr>
<td>FOS_199_SYC</td>
<td>9,287,614</td>
<td>99.81</td>
<td>93.39</td>
</tr>
<tr>
<td>FON_340_PHIL</td>
<td>10,038,683</td>
<td>99.73</td>
<td>92.48</td>
</tr>
<tr>
<td>FON_PDoug_PHIL</td>
<td>10,774,088</td>
<td>99.66</td>
<td>92.51</td>
</tr>
<tr>
<td>FOS_186_PHIL</td>
<td>10,258,077</td>
<td>99.78</td>
<td>91.86</td>
</tr>
<tr>
<td>FOS_188_PHIL1</td>
<td>8,866,016</td>
<td>99.69</td>
<td>91.92</td>
</tr>
<tr>
<td>FRS_234_POL</td>
<td>12,389,390</td>
<td>99.92</td>
<td>98.40</td>
</tr>
<tr>
<td>FRS_511_POL</td>
<td>11,552,960</td>
<td>99.83</td>
<td>92.87</td>
</tr>
<tr>
<td>FRN_364_POL</td>
<td>12,198,965</td>
<td>99.93</td>
<td>98.59</td>
</tr>
<tr>
<td>FRN_17_POL</td>
<td>5,981,049</td>
<td>99.94</td>
<td>98.39</td>
</tr>
<tr>
<td>FRSYD_1065_POL</td>
<td>9,903,314</td>
<td>99.93</td>
<td>97.68</td>
</tr>
<tr>
<td>FRSYD_2884_POL</td>
<td>12,712,178</td>
<td>99.92</td>
<td>98.55</td>
</tr>
<tr>
<td>FRN_274_SYC1</td>
<td>3,624,364</td>
<td>99.87</td>
<td>92.60</td>
</tr>
<tr>
<td>FRN_350_SYC</td>
<td>10,290,945</td>
<td>99.92</td>
<td>96.51</td>
</tr>
<tr>
<td>FRS_32_SYC</td>
<td>11,518,277</td>
<td>99.92</td>
<td>96.49</td>
</tr>
<tr>
<td>FRS_397_SYC</td>
<td>8,631,778</td>
<td>99.91</td>
<td>96.50</td>
</tr>
<tr>
<td>FRSYD_89_SYC</td>
<td>11,696,454</td>
<td>99.90</td>
<td>96.56</td>
</tr>
<tr>
<td>FRSYD_2873_SYC</td>
<td>9,193,680</td>
<td>99.88</td>
<td>96.50</td>
</tr>
<tr>
<td>FRN_332_SYC</td>
<td>11,025,824</td>
<td>99.86</td>
<td>95.50</td>
</tr>
<tr>
<td>FRS_238_SYC</td>
<td>11,776,247</td>
<td>99.88</td>
<td>95.57</td>
</tr>
<tr>
<td>FRSYD_1941_SYC</td>
<td>9,788,463</td>
<td>99.87</td>
<td>95.48</td>
</tr>
<tr>
<td>FRS_508_SYC</td>
<td>10,663,694</td>
<td>99.88</td>
<td>95.61</td>
</tr>
<tr>
<td>FRN_407_PHIL</td>
<td>9,946,326</td>
<td>99.79</td>
<td>90.16</td>
</tr>
<tr>
<td>FRN_310_PHIL</td>
<td>9,395,696</td>
<td>99.83</td>
<td>89.92</td>
</tr>
<tr>
<td>FRS_247_PHIL</td>
<td>10,963,241</td>
<td>99.79</td>
<td>90.17</td>
</tr>
<tr>
<td>FRS_40_PHIL</td>
<td>7,963,263</td>
<td>99.79</td>
<td>90.14</td>
</tr>
<tr>
<td>FRSYD_1822_PHIL</td>
<td>4,749,055</td>
<td>99.74</td>
<td>88.84</td>
</tr>
<tr>
<td>FRSYD_165_PHIL</td>
<td>8,222,265</td>
<td>99.82</td>
<td>90.15</td>
</tr>
<tr>
<td>FRN_299_WAT</td>
<td>10,388,116</td>
<td>99.92</td>
<td>95.09</td>
</tr>
<tr>
<td>FRN_337_WAT</td>
<td>9,550,950</td>
<td>99.91</td>
<td>95.12</td>
</tr>
<tr>
<td>FRS_514_WAT</td>
<td>9,524,174</td>
<td>99.94</td>
<td>95.94</td>
</tr>
<tr>
<td>FRS_239_WAT</td>
<td>10,456,714</td>
<td>99.93</td>
<td>95.97</td>
</tr>
<tr>
<td>Individual ID</td>
<td>Total number of reads (including both of the pair)</td>
<td>Percentage of reads aligned (including both of the pair)</td>
<td>Reads that mapped with Q20 or higher as percentage of all aligned reads*</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>FRSYD_567_WAT</td>
<td>13,453,889</td>
<td>99.93</td>
<td>95.89</td>
</tr>
<tr>
<td>FRSYD_1905_WAT1</td>
<td>12,324,737</td>
<td>99.94</td>
<td>96.06</td>
</tr>
<tr>
<td>FRN_372 HERO</td>
<td>6,977,488</td>
<td>99.85</td>
<td>92.77</td>
</tr>
<tr>
<td>FRN_349 HERO</td>
<td>9,068,864</td>
<td>99.88</td>
<td>92.95</td>
</tr>
<tr>
<td>FRS_517 HERO</td>
<td>9,024,466</td>
<td>99.87</td>
<td>92.94</td>
</tr>
<tr>
<td>FRS_244 HERO</td>
<td>12,394,088</td>
<td>99.84</td>
<td>92.84</td>
</tr>
<tr>
<td>FRSYD_1964_HERO</td>
<td>11,618,910</td>
<td>99.86</td>
<td>92.92</td>
</tr>
<tr>
<td>FRSYD_293_HERO</td>
<td>11,399,712</td>
<td>99.88</td>
<td>92.98</td>
</tr>
<tr>
<td>FRN_310 EUKO</td>
<td>9,632,111</td>
<td>99.78</td>
<td>93.19</td>
</tr>
<tr>
<td>FRN_406 EUKO</td>
<td>6,905,290</td>
<td>99.84</td>
<td>93.26</td>
</tr>
<tr>
<td>FRS_41 EUKO</td>
<td>7,973,427</td>
<td>99.83</td>
<td>92.77</td>
</tr>
<tr>
<td>FRS_239 EUKO</td>
<td>8,426,474</td>
<td>99.86</td>
<td>92.84</td>
</tr>
<tr>
<td>FWN_01-8 POL</td>
<td>17,947,652</td>
<td>99.98</td>
<td>99.73</td>
</tr>
<tr>
<td>FWN_161 POL</td>
<td>22,452,403</td>
<td>99.97</td>
<td>99.71</td>
</tr>
<tr>
<td>FWS_03-250 POL</td>
<td>10,994,084</td>
<td>99.93</td>
<td>99.66</td>
</tr>
<tr>
<td>FWS_137 POL</td>
<td>16,653,787</td>
<td>99.97</td>
<td>99.74</td>
</tr>
</tbody>
</table>

Table 2.8: A selection of mapping statistics per individual following mapping of the paired reads to each species meta-assembly. *A mapping quality score of 20 (Q20) means the aligner has designated the read as having a 1 in 100 chance of being misaligned.
2.4.4 Phylogenetic analysis of concatenated genome-wide SNPs across individuals within species

ML phylogenies were produced using concatenated alignments of SNPs from all sampled loci for each species. The trees were made per species and not per genus (as the cyt b phylogenies were) as the individuals from the different species within the same genus were aligned to different reference sequences. Also, it is important to note that these phylogenies are not gene trees but merely a convenient way to visualise the average genetic distance between individuals. They were produced to assess whether the expectations regarding the within species relationships, which were made following the barcoding are supported by these genome-wide summaries. The number of SNPs used for each species in these analyses is detailed in Table 2.9.
Under the assumption that each set of individuals were sampled from the same species (based on the mtDNA genealogies), the expectations regarding the average genetic distances are that: 1) branch lengths will be short between individuals from the same population (e.g. between the two North individuals) (i.e. external branches), and 2) branch lengths will be longer than those seen between individuals of the same population, when connecting individuals from different populations (e.g. between the North and South populations) (i.e. internal branches). If the branch lengths between all individuals are the same length (i.e. a star shaped tree), this could be indicative of a lack of population structure within the species (i.e. it forms one large continuous population).

Six species (P. nigriventris (ex F. watkinsiana), P. imperialis (ex F. rubiginosa), Sycoscapter short (ex F. rubiginosa), Watshamiella (ex F. rubiginosa), Eukobelea (ex F. rubiginosa) and Philotrepsis (ex F. obliqua)) out of eleven conform to the two expectations above for their North and South individuals (i.e. short within population branches separated by a longer between population branch) (Figure 2.17). The two of these that were also sampled from Sydney (P. imperialis (ex F. rubiginosa) and Watshamiella (ex F. rubiginosa)) suggest a lack of population structure between the South and Sydney individuals (i.e. the branch lengths within and between them are the same). The comparison of average genetic distances for five of these six species match the population structure inferred from the mtDNA genealogies. However the cyt b phylogeny for Watshamiella (ex F rubiginosa) (Figure 2.10) shows no structure between North, South or Sydney individuals, which is in contrast to its genome-wide SNP tree (Figure 2.17) that shows a long internal branch (relative to the within population branches) between the North and South/Sydney individuals.
Figure 2.17: Maximum likelihood phylogenies of concatenated genome-wide SNPs across individuals within species (Sycoscapter short (ex F. rubiginosa), Eukobelea (ex F. rubiginosa), Philotrypesis (ex F. obliqua), Pleistodontes nigriventris (ex F. watkinsiana) (top row) and Pleistodontes imperialis (ex F. rubiginosa), Watshamiella (ex F. rubiginosa) (bottom row)).

Two species (Sycoscapter long (ex F rubiginosa) and Herodotia (ex F rubiginosa)) out of the eleven show star shaped trees (i.e. the within and between population branch lengths are very similar) (Figure 2.18). The result in Sycoscapter long complements
the lack of structure between the individuals of each population presented in the cytb phylogeny (Figure 2.8). However the cytb phylogeny for *Herodotia* (Figure 2.11) showed two clades (albeit poorly supported ones) no evidence of which is apparent in these average genetic distance comparisons.

A further two species (*Pleistodontes* (ex *F. obliqua*) and *Sycoscapter* (ex *F. obliqua*)) out of the eleven did not show a pattern that conformed to the two expectations above. They both show (Figure 2.19) short external branches between the two South individuals and a short internal branch between the South and North pairs. The external branches leading to the two North individuals however show a large discrepancy in length. The large genetic distance, compared to all the other branch lengths, leading to a single North individual in both species suggests the two North individuals are not from the same population. The two North individuals that resulted in these long branches (one from *Pleistodontes* (ex *F. obliqua*) and one from *Sycoscapter* (ex *F. obliqua*)) were chosen for practical reasons (i.e. timing and financial considerations) over their placement in the mtDNA genealogies and this result.
Population genomics of a fig wasp community

confirms that in these cases the cytb phylogeny was successful in grouping these individuals with the South samples or in a separate clade for *Pleistodontes* (*ex F. obliqua*) and *Sycoscapter* (*ex F. obliqua*) respectively.

![Figure 2.19: Maximum likelihood phylogenies of concatenated genome-wide SNPs across individuals within species (Pleistodontes (*ex F. obliqua*), Sycoscapter (*ex F. obliqua*)).](image)

The last species (*Philotrypesis* (*ex F. rubiginosa*)) conforms to the two expectations above for its North and two of its South individuals (FRS_40_PHIL and FRS_247_PHIL) (Figure 2.20). However the other two South individuals cluster with one of the Sydney individuals (all showing external branches of the same length) whilst the second Sydney individual forms a separate cluster that is separated from its pair by a long internal branch. This result complements the cytb phylogeny (Figure 2.9) exactly in that the two Sydney individuals were taken from different clades (3 and 4), again in the end being chosen for practical reasons over their placement in the tree.
Figure 2.20: Maximum likelihood phylogeny of concatenated genome-wide SNPs across individuals within species *(Philotrypesis (ex F. rubiginosa)).*
2.5 Discussion

The results show that despite the difficulties of working with such small and often cryptic taxa it is possible to generate large individual genomic datasets, which provide thousands of polymorphic loci amenable to analysis for population genomic inference. Despite the manufacturers recommendations, I have successfully created these datasets using smaller than recommended DNA quantities that often had poor library fragment distributions. The data generated here provide a unique community level genomic dataset, which includes species that have never been studied in such detail before.

2.5.1 The utility of DNA barcoding

Given the requirement of identifying male fig wasps prior to use in whole genome library preparations the DNA barcode-based identification of genetic groups worked well. As outlined above and as carried out for most species, ideally we would verify species identity and population membership using their placement within clades of the cyt b phylogenies, which in turn were identified using sequences from previously identified samples. In a few cases, the choice of individuals was preferentially based on resource and time limitations over the certainty of their placements within the cyt b phylogenies.

The use of mtDNA barcoding is prevalent in species delimitation and phylogeographic studies. The mtDNA marker CO1 has been suggested as a universal barcode marker for all metazoans (Hebert et al., 2003), however it can be difficult to amplify in fig wasps (Darwell, 2013) meaning cyt b was the most appropriate marker for use in this study. Its use here also enabled the current sequences to be combined with those from other projects. The IDs of these previously generated sequences and the consequent clade groupings encompassing them were used, in this case, as the predominant basis for identification of the species sampled.

Despite its widespread use, there are several concerns regarding the use of mtDNA for species delimitation that need to be considered when using it in such analyses. The incongruence between gene and species trees due to incomplete lineage sorting can result in the misidentification of species groups (Degnan and Rosenberg, 2009). For example, mitochondrial divergence could be inferred between individuals of the same species if populations of the species had been split for a short period of time (on an evolutionary timescale) but come into secondary contact prior to being...
sampled (Toews and Brelsford, 2012). Under this scenario, the divide in the species was not long enough for nuclear loci to sort completely between the two populations but due to the more rapid rate of evolution in mtDNA, divergence between mtDNA loci can be detected and can lead to the inference of a misleading population history (Zink and Barrowclough, 2008). The opposite situation (i.e. different species can be grouped together on the basis of a single mtDNA locus) can also occur due to hybridisation amongst species (Nicholls et al., 2012). Such conflicting patterns in mtDNA phylogenies are estimated to be very common in animal taxa (Funk and Omland, 2003) highlighting the reasons to treat species groupings based solely on mtDNA loci with caution.

A further complication of using mtDNA as a barcoding tool is the presence of inherited symbionts (e.g. Wolbachia) (Hurst and Jiggins, 2005), a phenomenon that is particularly prevalent in fig wasps (Haine and Cook, 2005). The presence of a Wolbachia infection can lead to lower or higher levels of variation within mitochondria, which can distort the boundary between intra- and inter- species relationships (Hurst and Jiggins, 2005). It has also been shown that incomplete sampling across sister species and across geographic space, as well as the presence of shared spatial structure across species can provide misleading barcoding results (Nicholls et al., 2012). For these reasons the sample choice for Pleistodontes (ex F. obliqua) and Sycoscapter (ex F. obliqua) could have yielded the desired North/South sampling strategy despite the misleading placement in the cytb phylogeny of two of the North individuals (one from each species respectively). For example, the Pleistodontes (ex F. obliqua) North individual that is grouped amongst individuals from the South (Figure 2.7, clade 6) could represent hybridisation between the populations resulting in shared mitochondrial haplotypes. The low sample numbers, the poor support for the clade containing the other sampled North individual and the presence of a sample from the North population (one of the additional sequences used to populate the tree) within this ‘southern’ clade also supported this possibility.

An interesting result that highlights the potentially misleading nature of single mtDNA genealogies is the lack of structure seen amongst the Watshamiella (ex F. rubiginosa) individuals from all three populations in the cytb phylogeny (Figure 2.10). This is not supported by the average genetic distance between the populations (inferred from the genome-wide SNP analysis) that shows a long branch (Figure 2.17) between the North and South/Sydney populations. The long internal branch between populations is indicative of substantial structure between them that is not
apparent from the cytb sampling. Such a result could represent hybridisation between the individuals of these populations or the presence of Wolbachia across the individuals, both of which could cause this lack of population structure in the mitochondria (Funk and Omland, 2003; Hurst and Jiggins, 2005; Toews and Brelsford, 2012).

Confidence in the ID of the additional sequences used to populate the trees was high as they were not purely based upon a single mtDNA marker, they involved morphological analyses and a multi locus approach that made use of a nuclear marker (Segar, 2011; Darwell, 2013). The barcoding was crucial to this project as distinguishing between species morphologically using male fig wasps is impossible in some taxa. The desire to preserve as much DNA from each individual as possible was the main reason for not sequencing a nuclear locus along with the mtDNA cytb locus and in this case it may not have increased sampling success as the other limiting factors (i.e. timing and financial considerations) would still have been present. Ultimately given the restrictions encountered here (i.e. resource limitations, sample numbers), the barcoding process was highly successful in aiding the selection of appropriate samples for high throughput sequencing.

### 2.5.2 Detecting and removing contaminant reads

The blobplots produced to assess contamination within the samples highlight that most individuals show very little, if any, contamination with non-target organisms. This is surprising given that fig wasps, along with their host figs, are known to harbour diverse microbial communities (Martinson et al., 2012; Niu et al., 2015). The approach adopted here to assess contamination was conservative in that it removed reads that mapped to contigs identified at the phylum level. It was not meant as an assessment of the microbial community present but merely as a way of removing reads that likely originated from non-target organisms that would adversely affect the final assembly and interfere with the downstream demographic analyses. Other more detailed and targeted approaches could potentially be used to assess the microbial community structure present in this dataset (e.g. Niu et al., 2015).

The endosymbiont Wolbachia is known to have especially high infection rates in fig wasps and it has been shown that infection status can differ within species (Haine and Cook, 2005). All the individuals screened from the two disjunct populations of P. nigriventris (ex F. watkinsiana) by Haine and Cook (2005) showed the same pattern of Wolbachia incidence, Wolbachia were present in the northern population but absent in the southern population. It is interesting to note that the number of reads that
mapped to Wolbachia-identified contigs was much higher in the two North population *P. nigriventris* (*ex F. watkinsiana*) individuals compared with the numbers in the two South population individuals (102,809 and 145,604 reads compared with 3,351 and 5,329 reads for the North and South, respectively). The proportion of reads with respect to the total number obtained for each individual showed the difference between the North and South populations was significant (calculated using the total number of contaminant reads and total number of non-contaminant reads for both the North and South populations and a binomial test for comparing two proportions (X-squared = 95159, df = 1, p < 2.2e-16)). Reads that mapped to Wolbachia made up over 97% of those that mapped to Proteobacteria-identified contigs in all individuals except FWS_03-250_POL, where only 24% of reads mapped to Wolbachia-identified contigs. Conversely, all of the *P. imperialis* (*ex F. rubiginosa*) individuals screened by Haine and Cook (2005) over a wide geographical range were infected with Wolbachia. Again, my results are consistent with this finding: all six of the *P. imperialis* (*ex F. rubiginosa*) individuals sampled here produced a noticeable number of reads that mapped to Wolbachia-identified contigs (between 0.4% and 0.78% of the total number of reads for each of the 6 individuals). This result shows that the high percentage (44.2%) of reads that mapped to Proteobacteria-identified contigs in individual FRN_J7_POL (Figure 2.14) originates from another source of bacterial contamination rather than a Wolbachia infection.

The BLAST results were also checked for the presence of any sequences originating from the fig trees however 5 of the 11 fig wasp species showed no hits to *Ficus* sequences and the other 6 showed non-significant hits covering less than 1700 bases. This suggests there was very little, if any, carry over of plant material when these wasps were sampled.

### 2.5.3 Why does coverage differ between individuals and species?

In an attempt to create datasets that on average showed a uniform sequencing depth (coverage) per individual, the amount of each library to add to the sequencing pool was calculated based upon the DNA quantity in each library prep and the average insert size of each library. As hypothesised in the objectives, the ‘aimed for’ coverage of ~6 fold was not achieved for any individual and the level varies across species (Table 2.9). There are several factors that could have contributed to this. The most obvious is that the expected number of reads from each lane was not achieved, ~400 million reads per lane were expected and ~345 million and ~398 million for the first and second lanes respectively were received. This fact will most likely have
contributed to the lower than expected coverage across all the individuals however several other factors could have contributed to the variation seen between individuals. First of all, I made the assumption that all the species being sequenced had the same genome size. This meant that each was added to the pool in equal ratios. The estimated genome size assumed was 300 Mb. This was based on the genome assemblies of *Ceratosolen solmsi* (a pollinating fig wasp of the Agaonidae family) and *Nasonia vitripennis* (a generalist parasitoid of the family Pteromalidae) the most closely related organisms to have their genomes sequenced, which are 278 Mb and 295 Mb respectively (Werren et al., 2010; Xiao et al., 2013). As all the species were treated equally the assumption was that the sequencing resources would be distributed between all individuals pooled on the lane equally. However, if the genome sizes do differ it could lead to increased or decreased sequencing depth for smaller and larger genomes respectively. The genome assembly sizes give a rough indication of genome size and these do differ between species (Table 2.6). The assembly sizes (span) range from ~233 Mb to ~447 Mb. Another factor that likely affected the overall coverage achieved is how the numbers used to calculate the concentration of each library solution were obtained. The molar concentration for each library was calculated using the DNA concentration as measured by fluorometer (Qubit), the molecular weight of DNA and the average fragment size as measured by the Bioanalyzer software. Under the assumption that all the libraries pooled showed the same shape of distribution this method should produce roughly the same number of reads per individual in the pool. However when libraries with differing size distributions are pooled together it can create discrepancy in the efficiency of the sequencing effort across individuals, as small fragments will preferentially cluster on the sequencing lane (Nextera Library Validation and Cluster Density Optimisation – Illumina Technical Note, 2014). This means that although each library was calculated as being of equal concentration, the small fragment libraries will cluster more efficiently than large fragment libraries resulting in discrepancies in the output for each. This can also be a problem on a per individual basis. For example, in the case of a double peaked distribution (e.g. Figure 2.13) the smaller fragments in the library will cluster preferentially over the larger fragments but the large fragments will have contributed to the measure of DNA quantity and average fragment size which has the effect that the concentration calculated will be overestimated and therefore potentially produce less reads than expected.
2.5.4 The meta-assembly approach

Obtaining a reference genome is commonly achieved by sequencing a single (or a couple) individual to high coverage and assembling the reads to form a de novo reference assembly as part of a population genomic analysis pipeline (Miller et al., 2012; Liu et al., 2014b; Meyer et al., 2015; Palkopoulou et al., 2015; Frantz et al., 2016). Alternatively, analyses use a genome already available that has been produced as part of a targeted genome sequencing project (Adams et al., 2000; International Human Genome Sequencing Consortium, 2001; Lindblad-Toh et al., 2005; Honeybee Genome Sequencing Consortium, 2006; Li et al., 2010; Werren et al., 2010; Scally et al., 2012; Xiao et al., 2013). The latter approach relies upon genome availability of the organism of interest and the former approach is expensive and generally limits studies to single species (Crawford and Lazzaro, 2012; Buerkle and Gompert, 2013; Huang et al., 2016). To enable a community level approach that compares a number of species, a different method was employed to maximise the taxa sampled within the available resources. The meta-assembly approach used here makes use of the low coverage population level sampling by combining individuals of the same species to create a de novo reference assembly for each species. This approach ultimately involves a trade off, it is a cheaper way of building a reference assembly, which in this case allowed resources to be distributed across larger numbers of individuals/species, but combining more than one individual into the same assembly can cause the assembly to be more fragmented due to the higher level of variation present. Because of this, the approach is only suitable when combining reads from closely related individuals. A strength of this meta-assembly approach however is that it helps to avoid the effects of reference bias (i.e. bias toward the individual used to build the assembly) when mapping reads back to the de novo assemblies (Hearn, 2014), which has been shown to be an issue (Degner et al., 2009). If present, reference bias can lead to an underrepresentation of the variation across the individuals mapped to the reference as reads containing the SNPs present in the reference sequence could map preferentially.

2.5.5 Does meta-assembly quality correlate with genetic diversity?

As these datasets are intended for making cross species comparisons it is important to consider whether it matters that species’ assemblies differ in quality (as quantified by N50 and CEGMA scores). If lower assembly quality correlated with genetic diversity, it could suggest the estimates of genetic diversity are subject to bias that was introduced through the assembly process. To test whether assemblies
are poorer in species with higher genetic diversity, I plotted the scaffold N50 of each assembly (Table 2.6) against the overall per site diversity (the average pairwise diversity across individuals for each species) (Figure 2.21). This shows that there is no significant relationship between the two ($r = -0.457$, $p = 0.158$). This suggests that despite the large variation seen across species in assembly quality, quality is not correlated with the level of diversity measured in these species.

![Figure 2.21: Scaffold N50 of each assembly plotted against the overall per site diversity. The shaded areas represent the 95% confidence intervals.](image)

Another measure of assembly quality is the CEGMA scores (Table 2.7). These assess genome completeness based on the percentage of core eukaryotic genes (CEGs) that are present in the assembly. This is a commonly used metric of assembly quality that has been employed to assess the completeness of four insect genomes that were sequenced to high coverage: the honeybee, *Apis mellifera*, the parasitoid wasp, *Nasonia vitripennis*, the pollinating fig wasp, *Ceratosolen solmsi*, and the carpenter ant, *Camponotus floridanus* (Honeybee Genome Sequencing Consortium, 2006; Bonasio et al., 2010; Werren et al., 2010; Xiao et al., 2013). The complete CEGMA scores from this study range from 70.97% to 95.56%, which compare reasonably with those from the high coverage insect genomes that range from 94%-100%. There is no
relationship between the genetic diversity of each species and the CEGMA scores of their respective assemblies. Had a negative correlation between CEGMA score and genetic diversity been observed, it could suggest that the higher scores are a result of highly conserved regions (i.e. areas of low diversity) preferentially assembling. This was not observed ($r=0.316$, $p=0.343$) (Figure 2.22).

![Figure 2.22](image)

**Figure 2.22**: Complete CEGMA scores per species plotted against the overall per site diversity. The shaded areas represent the 95% confidence intervals.

### 2.5.6 Conclusions
This chapter highlights that despite its limitations, the use of DNA barcoding in cases where other forms of identification are difficult or even impossible is extremely successful. It has also demonstrated the ability to extract enough DNA from individual tiny wasps in order to prepare whole genome sequencing libraries. The libraries were successfully sequenced on the Illumina HiSeq and have yielded large genome-wide datasets for eleven species, nine of which will go on to be used in a series of population genomic analyses.

The sampling and bioinformatic pipeline presented here represents an effective way of obtaining individual level (i.e. as opposed to pooling the DNA from several
Population genomics of a fig wasp community

individuals) genome-wide datasets from tiny insects that could be applied to a range of taxa with similar size and/or sampling limitations. From the experience gained here, I would suggest that in order to maximise the final output (i.e. the size/number of genome-wide datasets) particular attention should be paid to the sample selection process. Careful consideration of the optimum sampling strategy and resource partitioning at this stage will result in final datasets that represent highly usable and valuable resources, not only for the intended study but also for future studies involving the sampled taxa.
3 Chapter 3: Profiling the demographic history of *Pleistodontes nigriventris*, the pollinator of *Ficus watkinsiana*

3.1 Abstract

The east coast of Australia presents a fascinating latitudinal transect with which to assess patterns of lineage distribution and population divergence. The transect covers a vast distance (over 3000 km) and includes a range of habitats and climates. Pollinating fig wasps represent a useful study system to investigate such phylogeographical patterns, particularly because of their dependence upon fig tree hosts; while fig wasps are tiny, fig trees are often obvious in their environments. This makes obtaining distributional data and the collection of samples easier and provides an excellent platform for considering interactions across trophic levels. Figs are an extremely large plant genus (~750 species) and are a keystone resource to many frugivorous vertebrates across their pan-tropical distribution. The Australian endemic, *Ficus watkinsiana*, has never been studied in detail and the relationship between its two distinct (North and South) populations of pollinating fig wasp, *Pleistodontes nigriventris*, has never been investigated before. This study uses low coverage genome-wide data to infer the demographic history of *P. nigriventris* and postulates the possible drivers of this history. *P. nigriventris* was found to show a Pleistocene divergence time between its North and South populations with the model significantly supporting the presence of ongoing post divergence gene flow, predominantly from the South into the North population. The climate oscillations of the Pleistocene had dramatic effects upon the Australian flora and fauna, despite glaciation not covering the continent widely. Intense drying during glacial periods caused rainforest contractions along the east coast, which likely contributed to the current distribution of *F. watkinsiana* and hence to the divergence history of its pollinator, *P. nigriventris*.
3.2 Introduction

3.2.1 Habitat barriers and dispersal

The genetic diversity present in extant individuals and populations reflects the evolutionary history of the species to which they belong. Climatic changes in particular have often been major factors in shaping evolutionary histories by influencing the size and connectedness of populations. These aspects of the demographic history of the population to which an individual belongs can be inferred from an organism’s genome. Climate oscillations experienced during the Plio-Pleistocene epochs caused widespread glaciation across the northern hemisphere forcing taxa into southern refugial areas followed by subsequent range expansions northwards during the warmer interglacial periods (Hewitt, 1996). In Australia, although widespread glaciation did not occur during these epochs, the environment was nevertheless severely affected by the climate cycles (Hewitt, 1996; Hewitt, 2000). They led to repeated expansion and contraction of mesic habitats in the Miocene, leading to an increase in arid zones and a decrease in rainforest along the humid regions of the east coast (Pepper et al., 2014). These changes in habitat created barriers to, and corridors for, gene flow in many taxa. The evolutionary consequences of these habitat fragmentations and range expansions will depend upon the length of time such breaks/connections in population continuity occurred for and upon the characteristics of the taxa affected, e.g. dispersal ability, generation time and population size. If populations are separated for long periods of time, they begin to diverge, a process that can ultimately lead to speciation (Hey, 2010). The process of divergence may or may not be accompanied by on-going gene exchange between the populations. Gene exchange could be an on-going process following the population separation, e.g. if the taxa involved can disperse effectively across the break, or it could occur during discrete time intervals, e.g. following complete isolation between two populations, local conditions could then change to allow the taxa to disperse across the break but only for a limited time. A species history is likely complex and could include a combination of divergence, expansion and/or gene exchange events. Here I will refer to the first scenario as the IM, isolation with migration, model (i.e. divergence with continuous post-divergence migration) and the second as the ADM model (i.e. divergence with a discrete instantaneous post-divergence admixture event). These simplified histories can be inferred from patterns of genetic diversity in extant populations (Nielsen and Slatkin, 2013). Population level inferences make important contributions to our understanding of how species have been affected by past environmental changes and hence can
provide an indication of how they may cope with future conditions (Hickerson et al., 2010).

### 3.2.2 Australian fig wasp phylogeography

Phylogeography is the study of the genealogical relationships between species/populations in geographic space (Hickerson et al., 2010). The term was coined to describe a discipline that would bridge the gap between classical population genetics and phylogenetics (Avise et al., 1987). The diverse array of wasp inhabitants associated with figs provides a model ecological community. They comprise many different species over multiple trophic levels, all of which are conveniently located in discretely samplable fig microhabitats (Segar et al., 2014). These features make them excellent study systems with which to assess phylogeographic patterns. Most phylogeographic studies involving the intimate relationship between fig trees and fig wasps have focussed on African, Asian and American species (Compton, 1993; Compton, 1994; West and Herre, 1994; West et al., 1996; Compton et al., 2009; Kobmoo et al., 2010; Chen et al., 2012; Cornille et al., 2012; Segar et al., 2013; Vieira et al., 2015). Few studies focus on endemic Australian figs and these have investigated only a minority of the 23 section *Malvanthera* species (Ronsted et al., 2008b; Segar et al., 2014). Given the tight specificity seen between pollinating fig wasps and their host figs (Herre et al., 2008; Cruaud et al., 2012), an interesting question to ask is whether the two trophic levels show similar patterns of genetic structure throughout their range (Wang et al., 2013; Liu et al., 2015; Wachi et al., 2016). It has been hypothesised that fig tree populations distributed over large distances are likely to have experienced habitat fragmentation in their histories and will therefore show a high degree of differentiation in both tree and wasp species (Wang et al., 2013). This is mainly attributed to the fact that the tiny, short lived fig wasps are poor fliers, potentially limiting gene flow in both the wasp and the tree through pollen transfer (Harrison, 2003; Wang et al., 2013). However, there is evidence that despite being poor flyers, fig wasps can disperse great distances (>160 km) using wind currents above the tree canopies (Ahmed et al., 2009). This ability is thought to be more pronounced in fig wasp species associated with monoecious trees as dioecious trees do not tend to grow as tall and therefore the wasps associated with them do not reach the heights where the wind currents are strongest (Harrison and Rasplus, 2006; Sutton et al., 2016).

Kobmoo et al (2010) used microsatellites to assess the patterns of genetic differentiation in the pollinator (*Ceratosolen fusciceps*) of a monoecious fig tree, *Ficus*
Population genomics of a fig wasp community

*racemosa.* *F. racemosa* is distributed from India to Australia and in this study they sampled from five sites, one located in south China and four across Thailand. They also included a sample from a single tree in Port Douglas, Australia. Despite the five south-east Asian sites spanning a distance of ~1600 km they found little genetic differentiation between any of the pairwise comparisons considered (all pairwise Fst values < 0.05). They found significant differentiation (all pairwise Fst values > 0.4) when comparing the south-east Asian populations with the Australian population however, suggesting possible on-going allopatric speciation, with more extensive sampling of the areas between the two populations required to confirm this. The authors conclude that high levels of gene flow across the Asian mainland sites is maintaining genetic homogeneity amongst these populations but that potentially, this level cannot be maintained across the expanses of ocean between mainland Asia and Australia (Kobmoo et al., 2010). This result was replicated in a study that used microsatellites to assess the genetic structuring found in an Asian dioecious fig, *Ficus pumila,* and its pollinator, *Wiebesia pumilae* (Liu et al., 2015). Both were sampled from 17 populations across ~1100 km of south China. The microsatellite markers showed a greater degree of differentiation in the host tree compared to its pollinator (Fst values 0.123 and 0.059 respectively). The authors conclude that long distance dispersal is maintaining the lack of variation between sites in the pollinator and, despite this being an expected characteristic of pollinators of monoecious species, the observation that *F. pumila* is a creeping fig and therefore capable of reaching the top of the canopy could potentially allow the wasps to disperse in the higher, stronger winds (Liu et al., 2015).

The most extensively studied Australian endemic fig species is *Ficus rubiginosa.* The pollinator of *F. rubiginosa,* *Pleistodontes imperialis,* has been investigated in detail using a number of mitochondrial and nuclear markers across its east coast distribution (Haine et al., 2006; Darwell et al., 2014; Sutton et al., 2016). One recent study found evidence that *P. imperialis* is actually a complex of five species (currently designated species 1-5) distributed sympatrically throughout the range of *F. rubiginosa* (Darwell et al., 2014). They used two mtDNA (cytb and CO1) markers and one nuclear marker (ITS2) and both distance and phylogeny based methods to assess the diversity of pollinators found on *F. rubiginosa.* The different markers and methods used all supported the presence of five sister species but there was a discrepancy in the level of substructure seen when using the mtDNA verses the nuclear markers (Darwell et al., 2014). Following from this study, Sutton et al (2016) presented the first comparative study of the genetic structures of an Australian fig...
wasp and one of its parasitoids across the majority of its natural range. The study used 6-9 microsatellite loci to assess differentiation between populations of *P. imperialis* (sp. 1) and its parasitoid *Sycoscapter long*. They found significant differentiation (pairwise Fst values 0.416-0.501) between the southerly and most northerly population of the pollinator, but no structure in the parasitoid across the entire range (all pairwise Fst values < 0.05), despite the study area spanning several known habitat barriers. The authors conclude that the dispersal ability of the parasitoid must be as good if not better than that of the pollinator over large distances (~2000 km sampled range) (Sutton et al., 2016).

Phylogeographic studies of fig wasp systems have primarily used a handful of either microsatellite or mtDNA/nuclear markers to infer genetic differentiation, and studies that scale up to thousands of genome-wide markers are scarce (see Wachi et al (2016), for a study that used ddRAD-seq (double digest restriction associated DNA sequencing) data to assess the genetic structure of four pollinating fig wasps in Japan and Taiwan). No population genomic studies of genetic structuring in Australian fig wasp community members have yet been attempted. This is surprising given the strong climatic gradient presented by this large latitudinal transect. This omission limits the ability to compare patterns across the pan-tropical distribution of *Ficus*. No studies, genomic or otherwise, have investigated the population genetic structure of the pollinator of the Australian endemic fig, *F. watkinsiana*, *Pleistodontes nigriventris*.

### 3.2.2.1 *Ficus watkinsiana* and its pollinator, *Pleistodontes nigriventris*

*Ficus watkinsiana* is a strangler fig endemic to Australia (Dixon, 2003; Male and Roberts, 2005) (Figure 1.8, Chapter 1). It is a member of the series *Rubiginosae* (section *Malvanthera*, subgenus *Urostigma*) (Ronsted et al., 2008b) and is currently recorded as being pollinated by a single fig wasp species, *Pleistodontes nigriventris* (subfamily *Agaoninae*, family *Agaonidae*) (Lopez-Vaamonde et al., 2002). *Ficus watkinsiana* is a large hemi-epiphytic species found in two distinct populations of upland rainforest, one in north-east Queensland and the other in south-east Queensland and north-east New South Wales (Dixon, 2003) (Figure 1.7, Chapter 1). These populations are around 1300 km apart, the majority of the intervening habitat being unsuitable dry forests (Haine and Cook, 2005). *F. watkinsiana* is often found above the forest canopy and, like the majority of monoecious figs, different trees produce fruit asynchronously throughout the year (Dixon, 2003). The trees generally produce small crops (~10-2000) of their large figs (Figure 1.8, Chapter 1) but an
Population genomics of a fig wasp community

individual tree can produce fruit for up to 5 months of the year (Dixon, 2003). *Ficus watkinsiana* is a monoecious species. Generally, monoecious fig species are larger than the small, shrubby dioecious species, allowing their pollinating wasps to reach greater heights and therefore the stronger winds found above forest canopies (Ahmed et al., 2009; Yang et al., 2015). A large area of unfavourable habitat that includes two known biogeographic barriers, the Burdekin Gap and the St. Lawrence Gap (detailed below), splits the two current *F. watkinsiana* populations (Figure 1.3, Chapter 1). The trees are sometimes planted in ornamental gardens and so it is possible that there are trees found between the two populations in parks or city gardens but certainly not in great numbers (J. Cook, personal communication). Despite the long distance dispersal capacity of pollinators, the ~1300 km gap and the rarity of ‘stepping stone’ *F. watkinsiana* trees within the gap suggest we may see significant genetic differentiation due to very low or non-existent gene flow between these pollinator populations.

3.2.3 The Burdekin and St. Lawrence Gaps

Although widespread glaciation did not occur in Australia, the climatic oscillations during the Pleistocene had a significant impact upon its vegetation cover (Barrows et al., 2002). The cooling temperatures reduced rainforest cover, restricting them to areas of higher rainfall along the eastern and southern coasts (Markgraf et al., 1995). This in turn made way for the extension of shrub and grasslands (Markgraf et al., 1995). Currently, the majority of the Australian continent is dominated by arid or semi-arid vegetation (Byrne et al., 2008), the key exception being the eastern coastal areas (Chapple et al., 2011b). The east coast of Australia is predominantly mesic habitat interspersed with areas of rainforest (Byrne et al., 2011) and harbours high organismal diversity (Chapple et al., 2011b). The maintenance of this mesic habitat is facilitated by the presence of the Great Dividing Range, a large mountain range that runs the length of the east coast (Chapple et al., 2011b). This latitudinal strip of habitat is subdivided by several known and potential biogeographic barriers, which previous studies have shown to have a major role in species divergence and ultimately speciation in the Australian flora and fauna (Chapple et al., 2011b; Bryant and Krosch, 2016). The Burdekin and St. Lawrence Gaps are two of the east coast’s most significant biogeographical barriers (Figure 1.3, Chapter 1). The Burdekin Gap is the largest dry land corridor on the east coast, located between the towns of Mackay and Townville. Around 350 km south is a second smaller lowland dry corridor, known as the St. Lawrence Gap (Bryant and Krosch, 2016). Palaeovegetation records for both are poorly characterised, meaning that accounts
of the history of their formation and stability are limited (Bryant and Krosch, 2016). Both, however, have been implicated as breaks in species distributions and drivers of divergence in a range of taxa (James and Moritz, 2000; Pope et al., 2001; Schäuble and Moritz, 2001; Nicholls and Austin, 2005; Brown et al., 2006; Dolman and Moritz, 2006; Baker et al., 2008; Edwards and Melville, 2010; Chapple et al., 2011b; MacQueen et al., 2012; Rix and Harvey, 2012; Burke et al., 2013; Smissen et al., 2013; Bryant and Fuller, 2014; Bryant and Krosch, 2016) (see Chapter 1).

3.2.4 Genomic scale demographic inference

Demographic inference has become an important tool for understanding the evolutionary past of an organism. Moreover, comparing such inferences across species contributes to the understanding of how natural communities are structured (Hickerson et al., 2010; Smith et al., 2011). Such inferences can be made using the sequence variation present in the DNA of extant individuals.

The complete ancestral history of a sample(s) of sequences is described by the ancestral recombination graph (ARG), which includes both coalescent and recombination events, where, looking backwards in time, a coalescence event occurs where two lineages join (i.e. share a common ancestor) and a recombination event occurs at the splitting of two lineages (i.e. different parts of the sequence are derived from different ancestors) (Hoban et al., 2012; Kelleher et al., 2016). Accounting for recombination in gene trees adds extra complexity to the inference process that is impossible to incorporate in current inference methods. Because of this, many inference methods assume full recombination between sites, i.e. they ignore linkage among sites and use only polymorphic sites with no consideration for the number of intervening invariant sites, hence each locus is assumed to be independent. The explosion in available sequence data over recent years has motivated the exploration of a wide range of demographic inference methods (Excoffier and Heckel, 2006; Pool et al., 2010; Schraiber and Akey, 2015). An important consideration for methods using such large datasets is computational tractability.

A common way to summarise whole genome datasets is via the site frequency spectrum (SFS). The SFS is the distribution of single nucleotide polymorphism (SNP) frequencies seen across the genome (or a section of it) in a single population sample or across samples from multiple populations (joint SFS) (Gutenkunst et al., 2009; Lukic and Hey, 2012; Excoffier et al., 2013; Xue and Hickerson, 2015). The expected SFS under a given population history can be approximated using coalescent simulations (Excoffier et al., 2013) or using the numerical solution to a
diffusion equation (Gutenkunst et al., 2009). As the SFS is based on SNPs, inference methods based on the SFS ignore linkage between loci and treat sites as independent.

The biggest drawback of SFS methods in general is that they ignore linkage information between variable sites (Theoretical Evolutionary Genetics, J. Felsenstein, http://evolution.genetics.washington.edu/pgbook/). There have been attempts to approximate the coalescent model to incorporate recombination to make use of linkage information in sequence data whilst maintaining computational efficiency. A version of this approximated model is the sequential Markov coalescent (SMC) proposed by McVean and Cardin (2005). They model the non-Markovian coalescent with recombination process as a Markovian process along the genome. Thus, the SMC assumes that each recombination event depends only on the previous one, hence massively reducing the state space required for computing likelihoods. It has been shown that the SMC is a good approximation of the coalescent with recombination model for the purpose of inference (McVean and Cardin, 2005).

3.2.5 Exploring ancestral population size changes using PSMC

Since its development, different methods have used the SMC to infer demography (Li and Durbin, 2011; Harris and Nielsen, 2013; Schiffels and Durbin, 2014) and additional modifications have been made to improve its agreement with the full model (the ARG) (Marjoram and Wall, 2006). The pairwise sequentially Markovian coalescent (PSMC) developed by Li and Durbin (2011) infers the historical changes in effective population size through time from a pair of homologous DNA sequences, most commonly from a single diploid individual. It uses the density of heterozygous sites to infer each local genealogy from which the distribution of time to the most recent common ancestor is estimated. The rate of coalescence is inversely proportional to the effective population size ($N_e$), allowing the ancestral $N_e$ trajectory for the pair of sequences to be inferred. This method has been applied to a wide range of taxa, e.g. dogs, polar bears, pigs, bumblebees, apes and parasites (Groenen et al., 2012; Miller et al., 2012; Prado-Martinez et al., 2013; Freedman et al., 2014; Wallberg et al., 2014; Chen et al., 2016; Small et al., 2016). The main drawback is that because the model only makes use of a single diploid individual it cannot infer recent $N_e$ changes due to a lack of recent coalescent events. For example, in humans it is unable to infer $N_e$ changes less than ~20,000 years ago (Li and Durbin,
This particular limitation has been addressed by Schiffels and Durbin (2014) in their framework, the multiple sequentially Markovian coalescent (MSMC). MSMC can use up to four diploid individuals to infer changes in ancestral Ne over time from as recent as 70 generations ago (~2,000 years ago in humans) (Schiffels and Durbin, 2014). However, these genomes need to be phased, which is not the case for a PSMC analysis (Schiffels and Durbin, 2014).

3.2.6 Demographic inference using the blockwise SFS likelihood method

An important and widespread assumption across demographic inference methods is that there is no recombination within loci and free recombination between loci. This allows each locus to be treated as an independent replicate of the coalescent, simplifying likelihood calculations. Lohse et al (2016) have implemented a likelihood approach that uses short stretches of aligned sequence (termed sequence blocks) to fit divergence models and infer population histories. Assuming a sample of four haploid individuals, two each from two populations, there are 15 possible labelled rooted genealogies (Felsenstein, 2003). This is reduced to six when removing phase (i.e. removing labels that distinguish the within population individuals) and further reduced to two genealogies in the absence of root information (Felsenstein, 2003). Under this ‘two by two’ sampling scheme each sequence block can be summarised by counting four possible mutation types: heterozygous sites in population A (Ahet), heterozygous sites in population B (Bhet), fixed differences between populations A and B (Fixed), and shared heterozygous sites between populations A and B (Shared) (Lohse et al., 2016). This blockwise summary is known as the blockwise SFS (bSFS) (Figure 3.1). Heterozygous sites unique to each population contribute information on within population diversity and allow effective population sizes to be estimated for each population. They also contribute to post-divergence gene flow estimates. For example, unidirectional gene flow is anticipated to increase the heterozygosity of the receiving population as new alleles join the gene pool. The relative number of fixed and shared sites between populations provides information on how divergent the two populations are. A recently (on the coalescent timescale) diverged population would be expected to share a higher proportion of heterozygous sites due to incomplete lineage sorting of ancestral polymorphisms. An old divergence event however, is expected to be characterised by a higher proportion of fixed differences. This is due to one or other allele being lost from each lineage over time through genetic drift (i.e. complete lineage sorting). A high rate of gene flow is
expected to increase the proportion of shared heterozygous sites due to the ongoing mixing of the two populations’ gene pools.

Figure 3.1: Example bSFS configurations for a sample of four haploid individuals, two from population A and two from population B. The possible mutation types (for which counts are given in a vector $k$, with 4 values by each diagram) are, in turn: (a) heterozygous sites in population A (Ahet), (b) heterozygous sites in population B (Bhet), (c) fixed differences between populations A and B (Fixed), and (d) shared heterozygous sites between populations A and B (Shared). The red and blue lines show the differences between site types and the purple dots represent the mutations that have occurred to produce the different site types.

The choice of block lengths to use in such an analysis needs to strike a balance between being long enough to include an average of more than one polymorphic site per block, but short enough to assume that little or no within-block recombination has occurred. The power to infer different demographic scenarios increases with the number of blocks (the number of loci) and (in contrast to some other approaches) this method can be applied using an infinitely large number of blocks. Lohse et al (2016) use the generating function of genealogical branches to calculate the probability of seeing different blockwise mutational configurations.
under the specified model. The joint likelihood of the model is the sum of log likelihoods across all blocks. The joint likelihood is maximised for the model to find maximum likelihood estimates of the model parameters. In theory, this method can be used to model any set of population histories. In practice, it is limited to a small set (~3) of interacting populations. It is able to model both types of gene exchange that I am interested in: divergence with continuous migration (IM) and divergence with instantaneous admixture (ADM) (Figure 3.2).

![Figure 3.2: Schematics of the two divergence models, IM (divergence with continuous migration) and ADM (divergence with instantaneous admixture). Gene flow can be modelled in either direction and each population can have a different Ne.](image)

In common with other demographic inference methods based on the bSFS, the likelihood calculations assume the infinite sites model of molecular evolution, i.e., that there are an infinite number of sites at which mutations can occur, so no site mutates more than once, and no intralocus (i.e. within block) and free interlocus (i.e. between block) recombination. The likelihood calculation excludes blocks that contain both Fixed and Shared sites, as these would represent a violation of either the assumption of no within block recombination or the infinite sites assumption, given that this configuration could only be generated by past recombination.

Nurnberger et al (2016) applied this method to a transcriptome dataset from two hybridising species of fire-bellied toads in Europe. The study used a single diploid individual from each species to create a dataset of just under 1500 orthologous
sequence blocks. The authors summarised the datasets in two ways, both of which were used to fit models of divergence: 1) the SFS (randomly sampling one variable site per block), 2) the bSFS (as described above). The bSFS provides more information from linked polymorphisms, enabling more complex models to be fitted, but is more likely to be affected by biases introduced if the model assumptions are violated. The authors fitted divergence models without (null model) and with unidirectional (in both directions) gene flow using both datasets. They found that despite a known ancient divergence time, the likelihood analyses were able to detect on going gene flow between the two species albeit at a very low level. They also report that overall, both datasets gave similar point estimates (Maximum Likelihood Estimates (MLEs) of the model parameters) with the exception that the bSFS data provided the resolution needed to infer the presence of different effective population sizes in the two populations (Nurnberger et al., 2016).

### 3.2.7 Objectives

The overall aim of this chapter is to assess the population history of *P. nigriventris* using a whole genome dataset created from four haploid male wasps, two from each population (sample collection, laboratory and initial bioinformatics protocols are described in Chapter 2). Demographic inferences were made by fitting a series of divergence models using the novel bSFS likelihood method and via the pairwise sequentially Markovian coalescent (PSMC) approach. Within the overall objective, this chapter aims to answer the following more specific questions: 1) Is there evidence of genetic divergence between these two populations of pollinating fig wasp, and if so, when did it occur? 2) Does the time of divergence support studies in other taxa that have inferred divergence as a result of the biogeographic barriers found between the two populations? 3) Is there a signal of gene flow having occurred since divergence, and if so, which direction predominates? 4) If present, does the gene flow signal in the data allow the most likely mode of gene flow (i.e. continuous post-divergence gene flow vs a single discrete admixture event) to be determined? And finally, 5) Is the inferred population history obtained from the bSFS likelihood method supported by the PSMC analysis? I outline my hypotheses regarding the population history of *P. nigriventris* below and briefly state the reasons behind them.

The host fig trees (*F. watkinsiana*) of *P. nigriventris* are rainforest specialists and are currently found distributed as two distinct populations on the east coast of Australia, one situated around Cairns and the other around Brisbane. The


populations are currently separated by a large expanse of unsuitable habitat. What is unknown is how long they have been separated by such a large distance. The climate cycles of the Pleistocene are known to have had severe impacts upon the distribution of the flora and fauna of Australia. The cold/dry glacial periods favoured dry forest and grass species whereas the warm/wet interglacials allowed rainforest taxa to expand. These climatic fluctuations caused the rainforests to become increasingly fragmented, which has left signals in the genetics of the organisms affected. For these reasons, I hypothesise the two *P. nigriventris* populations, along with their host trees, have become increasingly affected by the glacial cycles throughout the Pleistocene, gradually becoming further apart, resulting in a signal of significant divergence between them. Pollinating fig wasps are poor fliers but can travel large distances using the wind above the tree canopies. Despite this, the vast expanse of unsuitable habitat that currently separates the two *P. nigriventris* populations makes significant gene flow between them seen unlikely. However, rare but unsampled ‘stepping stone’ trees between the two populations may allow for low levels of dispersal. The fact that the fig wasps predominantly travel from tree to tree carried by the wind suggests the overriding wind direction may indicate which direction gene flow is most likely to occur. Although current records show the predominant direction to be South to North, not knowing how such weather patterns have changed over longer timescales makes it difficult to hypothesise a direction of gene flow based on wind direction. Overall I expect to find evidence of significant divergence between the two fig wasp populations, with the potential for limited gene flow.
3.3 Methods

3.3.1 Model fitting

VCF and BED files were generated through the pipeline outlined in Chapter 2. The VCF file contains all the filtered variant sites called across the four *Pleistodontes nigriventris* individuals. The BED file details the positions of the callable regions in each contig and the length of each region.

3.3.1.1 Block cutting

I cut the genomic data into a number of aligned sequence blocks of a fixed length. The blocks were cut from homologous regions aligned across all individuals per species. The average pairwise \( \pi \) across individuals was used to determine the length of these sequence blocks. I chose to aim for 1.5 variant sites per block. Dividing this by the average pairwise \( \pi \) estimate gave me a starting block length of 387 bases for *P. nigriventris*.

I prepared the sequence blocks using a custom python (Python 2) script. Firstly, the script requires a list of unique contig IDs found across all individuals. I created this list using *tabix* (version 1.2.1) with the ‘–l’ option. Secondly, the script uses the positions in the BED file to designate regions (of the specified block length) that are callable across contigs. Thirdly, it uses a number of criteria to filter the blocks: each block must be made up of sequence from the same contig, contigs are filtered so none that are less than two times the block length contribute to the blocks and blocks must be made from contig regions no longer than two times the block length. Fourthly, it summarises each block in terms of its absolute position on the contig, each variant site within the block and the site call for each individual at that variant site. Rarely, regions may occur that have been deemed callable by the *CallableLoci* function but contain sites that the *HaplotypeCaller* has not been able to confidently call as variant or not in one or more individual (‘None’ sites). Because an excess of these sites in a block makes it unreliable I removed all blocks with more than 5 ‘None’ sites. Lastly, the script further summarises the blocks so that each consists of a vector of the four mutational types (Ahets, Bhets, Fixed, Shared) possible when considering four haploid (or two diploid) individuals. The blocks that failed the 4-gamete test (i.e. those that include both fixed differences and shared heterozygous sites) were not included in the model inference. The vector summaries of the blocks were exported as a text file in a *Mathematica* (Wolfram Research, Inc., Mathematica, Version 10.4, Champaign, IL (2016)) readable format. The script relies upon *pyvcf* (version 0.6.7) and *pysam* (version 0.8.4) to process the VCF files.
3.3.1.2 Data summaries, k-max values and configuration counts

The text file containing the block vectors (i.e. the per block summaries of each mutation type (the bSFS, described above)) was read into a Mathematica (Wolfram Research, Inc., Mathematica, Version 10.4, Champaign, IL (2016)) notebook. An initial exploration of the data consisted of calculating several summaries within the Mathematica notebook (e.g. proportion of each mutation type (Ahet, Bhet, Fixed, Shared), average per site heterozygosities, \( Fst \) and number of four gamete violations). Here I refer to blocks consisting of the same set of values as sharing a single mutational configuration, whilst blocks with different values for each count in the vector represent different configurations. The number of alternative possible configurations increases rapidly with the number of alternative count values allowed (\( k\)-max values) for each mutational type (Figure 3.3). The majority of the possible configurations will not be represented in the data and the computational burden can be reduced by restricting the space of possible configurations whilst retaining the most informative (i.e. those that appear more than once). One way of doing this, used here, is to lump rarely observed counts above some threshold value (\( k\)-max) into a single category. For example, a \( k\)-max value of 2 specifies that there can be 4 possible counts for each mutation type per block; zero mutations, 1 mutation, 2 mutations or > 2 mutations. The \( k\)-max values across mutation types were chosen using the Mathematica function ‘FindkMax’ (written by K. Lohse). The function assesses what proportion of the data falls into the > \( k\)-max category under differing \( k\)-max values. Any configurations that fall into this category will not have an exact probability associated with them, rather a combination of marginal probabilities. The idea is to maximise the number of configurations with exact probabilities whilst maintaining computational efficiency. The vector of \( k\)-max values used here was \{2, 2, 3, 2\} and the counts of mutational configurations across the dataset summarised using the Mathematica function ‘configCounts’ (written by K. Lohse). The data were also summarised as frequencies of each mutation class across the entire dataset (rather than blockwise) by randomly sampling a single SNP per block, to generate a standard SFS dataset for assessing model fit.
3.3.1.3 Likelihood analyses of divergence models

The likelihood method described in Lohse et al (2016) and introduced above was used to fit a series of divergence scenarios. The IM model assumes that an ancestral population with a constant effective population size ($Na$) splits into two descendant populations (A and B), one that maintains the same ancestral population size $Na$, whilst the other is free to change to $(1/B) * Na$, (where $B$ is a scaling parameter) at time $T$ (scaled by $2Ne$). Continuous post-divergence unidirectional migration occurs at rate $M = 4Nem$ migrants per generation. A model was derived to represent migration in both directions (A->B and B->A). A->B refers to a lineage from population B coming from population A moving backwards in time (i.e. population B is receiving migrants from population A) and vice versa for the B->A model. To enable model comparison, simpler versions of these models, where for example, $M$ was set to zero or both populations were assumed to share the same $Na$, were also fitted.

The ADM model assumes the same divergence history as the IM model with the only difference being the way the mode of gene flow is modelled. A unidirectional admixture event between the two populations (A and B) is modelled to have occurred at time $T1$ (scaled by $2Ne$) going back in time from the present. This represents an instantaneous movement of a fraction ($f$) of lineages from one population into the other. A->B refers to the admixture fraction of lineages in population B coming from population A backwards in time (i.e. population B is
receiving a fraction of population A) and vice versa for the B->A model. The divergence event occurred at time $T_2$ before the admixture event ($T_1$) so the time of divergence is calculated as $T_1 + T_2$ (scaled by $2N_e$). Again simpler versions of these models were fitted for comparison. Details of the full range of models tested (for both the IM and ADM scenarios) are described in Tables 3.1 and 3.2 and represented diagrammatically in Figure 3.4.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$N_e$ symmetry</th>
<th>Migration direction*</th>
<th>Model parameters</th>
<th>Model shorthand*^</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strict divergence</td>
<td>Symmetric</td>
<td>M-&gt;0</td>
<td>$\theta, T$</td>
<td>STRDIV1Ne</td>
</tr>
<tr>
<td>Divergence with migration</td>
<td>Symmetric</td>
<td>A-&gt;B</td>
<td>$\theta, T, M$</td>
<td>IM1NeAB</td>
</tr>
<tr>
<td>Divergence with migration</td>
<td>Symmetric</td>
<td>B-&gt;A</td>
<td>$\theta, T, M$</td>
<td>IM1NeBA</td>
</tr>
<tr>
<td>Strict divergence</td>
<td>Asymmetric</td>
<td>M-&gt;0</td>
<td>$\theta, T, B$</td>
<td>STRDIV2Ne(a)</td>
</tr>
<tr>
<td>Strict divergence</td>
<td>Asymmetric</td>
<td>M-&gt;0</td>
<td>$\theta, T, B$</td>
<td>STRDIV2Ne(b)</td>
</tr>
<tr>
<td>Divergence with migration</td>
<td>Asymmetric</td>
<td>A-&gt;B</td>
<td>$\theta, T, B, M$</td>
<td>IM2NeAB(a)</td>
</tr>
<tr>
<td>Divergence with migration</td>
<td>Asymmetric</td>
<td>B-&gt;A</td>
<td>$\theta, T, B, M$</td>
<td>IM2NeBA(a)</td>
</tr>
<tr>
<td>Divergence with migration</td>
<td>Asymmetric</td>
<td>A-&gt;B</td>
<td>$\theta, T, B, M$</td>
<td>IM2NeAB(b)</td>
</tr>
<tr>
<td>Divergence with migration</td>
<td>Asymmetric</td>
<td>B-&gt;A</td>
<td>$\theta, T, B, M$</td>
<td>IM2NeBA(b)</td>
</tr>
</tbody>
</table>

Table 3.1: IM model descriptions. *Migration direction: A->B signifies gene flow occurs from population A into population B, B->A signifies gene flow occurs from population B into population A, M->0 signifies the model does not include migration. **^Meaning of the codes in the model shorthand: STRDIV = model with no gene flow, IM = model with unidirectional gene flow, 1Ne = both populations share the same $N_e$, 2Ne = each population is allowed to have a different $N_e$ (although both remain constant through time), AB = gene flow is modelled going from Population A into Population B, BA = gene flow is modelled going from Population B into Population A, (a) = Population A is scaled against the ancestral $N_e$ (in this case the $N_e$ of Population B), (b) = Population B is scaled against the ancestral $N_e$ (in this case the $N_e$ of Population A).
<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ne symmetry</th>
<th>Admixture direction*</th>
<th>Model parameters</th>
<th>Model shorthand*^</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strict divergence</td>
<td>Symmetric</td>
<td>T1→0 and f→0</td>
<td>θ, T2</td>
<td>STRDIV1Ne</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Symmetric</td>
<td>A→B (T1→0)</td>
<td>θ, T2, f</td>
<td>ADM1NeT10AB</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Symmetric</td>
<td>B→A (T1→0)</td>
<td>θ, T2, f</td>
<td>ADM1NeT10BA</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Symmetric</td>
<td>A→B</td>
<td>θ, T1, T2, f</td>
<td>ADM1NeAB</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Symmetric</td>
<td>B→A</td>
<td>θ, T1, T2, f</td>
<td>ADM1NeBA</td>
</tr>
<tr>
<td>Strict divergence</td>
<td>Asymmetric</td>
<td>T1→0 and f→0</td>
<td>θ, T2, B</td>
<td>STRDIV2Ne(a)</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>T1→0 and f→0</td>
<td>θ, T2, B</td>
<td>STRDIV2Ne(b)</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>A→B (T1→0)</td>
<td>θ, T2, B, f</td>
<td>ADM2NeT10AB</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>B→A (T1→0)</td>
<td>θ, T2, B, f</td>
<td>ADM2NeT10BA</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>A→B</td>
<td>θ, T1, T2, B, f</td>
<td>ADM2NeAB(a)</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>B→A</td>
<td>θ, T1, T2, B, f</td>
<td>ADM2NeBA(a)</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>A→B</td>
<td>θ, T1, T2, B, f</td>
<td>ADM2NeAB(b)</td>
</tr>
<tr>
<td>Divergence with admixture</td>
<td>Asymmetric</td>
<td>B→A</td>
<td>θ, T1, T2, B, f</td>
<td>ADM2NeBA(b)</td>
</tr>
</tbody>
</table>

Table 3.2: ADM model descriptions. *Admixture direction: A→B signifies gene flow occurs from population A into population B, B→A signifies gene flow occurs from population B into population A, f→0 signifies the admixture fraction is set to zero (i.e. no lineages move from one population into the other), T1→0 signifies the admixture time is set to zero (i.e. the admixture event effectively happened yesterday). *^Meaning of the codes in the model shorthand: STRDIV = model with no admixture event, ADM = model with a single instantaneous admixture event, 1Ne = both populations share the same Ne, 2Ne = each population is allowed to have a different Ne (although both remain constant through time), T10 = time of the admixture event has been set to zero, AB = admixture event is modelled going from Population A into Population B, BA = admixture event is modelled going from Population B into Population A, (a) = Population A is scaled against the ancestral Ne (in this case the Ne of Population B), (b) = Population B is scaled against the ancestral Ne (in this case the Ne of Population A).
3.3.1.4 Scaling parameter estimates

The bSFS scheme assumes the same mutation rate across all blocks. As there are currently no estimated mutation rates ($\mu$) for any species of fig wasp, to scale the inferred parameter estimates from the models, I have used the per generation *Drosophila melanogaster* mutation rate estimated by Keightley et al (2014) of $2.8\times10^{-9}$. To be able to convert time estimates from generations into years, we need to know the generation time for *P. nigriventris*. I have assumed four generations per year (J. Cook, personal communication). I will discuss the reasons for these estimates and any implications of possible inaccuracy in the Discussion.

The value of $\theta$ inferred by the models is a per block estimate and so is divided by the block length (387) to obtain a per site theta ($\theta_{bp}$) to use in the calculations. The $Na$ is calculated as $Na = \theta_{bp}/4\mu$. If the model includes asymmetric $Ne$ values, the scaled $Ne$ is calculated using $(1/B)*Na$. The divergence time (in generations) in the IM models is calculated as $T*2Ne$ and in the ADM models as $(T1+T2)*2Ne$ (the time of the admixture event is $T1*2Ne$). I have divided the divergence time in generations by 4 to give an estimate in years.
3.3.1.5 Recombination rate estimation

Currently there are no estimated recombination rates ($r$ the per generation recombination rate, $\rho$ ($4Ne$) the population recombination rate (Stumpf and McVean, 2003)) for any fig wasp species. An estimate is required to simulate data with recombination, which is used to obtain confidence intervals for the parameter estimates (see below). A recombination rate for *P. nigriventris* was estimated using a two-locus generating function that co-estimates $\rho$ (scaled by $2Ne$) and $\theta$. The equation for the model is derived in Lohse et al (2011) (equation 19) and it assumes a pairwise sample (here 2 haploid genomes) from a single panmictic population.

The data were prepared to be input into the model using an adapted version of the block cutting python script described above. The method requires a pairwise alignment so in this case I used the data from the two northern individuals (FWN_01-8_POL and FWN_161_POL). The script filters the contigs by length and retains the longest 10%. A given threshold specifies which of these contigs are retained based on the proportion of callable sites present in each contig. I used a threshold of 0.7, meaning that each contig could contain no more than 30% non-callable sites. The output consisted of two lists: one containing the contig names, the number of callable sites within it and its original length, the other containing the positions of the variant sites along the contigs. This output file is then read into a Mathematica notebook. Within the notebook the contigs are divided into blocks (387 bases long), the number of variable sites present in each one is counted and split into datasets that each contain pairs of blocks at varying distances apart. The Mathematica function ‘configCounts’ (section 3.3.1.2) was used to count how many of each paired configuration was present in each dataset. Plotting the $\rho$ estimates for these datasets showed that they become stable with increasing distance between the pairs. To estimate $r$, the $\rho$ estimates were multiplied by 2 (to scale by $4Ne$) and divided by the distance between the blocks (per block estimate) before being divided by the corresponding $\theta$ estimate and being multiplied by $\mu$ (as above) ($r=(\rho/\theta)\mu$). To assess which distance gave the best-supported estimates, I obtained the log likelihoods for a range of $\rho$ estimates (covering all those output by the model) by fixing $\theta$ and plotting the difference in log likelihoods (compared to the highest) for each distance. The likelihood surface gets steeper with increasing distance until a distance $\geq$10 blocks where the support remains stable. I used the estimate obtained from the dataset with pairs 10 blocks apart in all subsequent analyses that required an $r$ estimate.
3.3.1.6 **Bootstrapping – correction for LD**

The bootstrap analyses are required to test for the effects of linkage disequilibrium (LD) in the data as, despite the assumptions of independence amongst the blocks, undoubtedly a proportion of them will be from adjacent regions and therefore could be linked. Assuming each block is independent under unknown levels of linkage results in overconfidence in the estimates produced. The likelihood per block is multiplied across all blocks to give the model likelihood, which therefore increases linearly with the number of blocks. Given that some of the blocks will be highly correlated this results in a highly inflated likelihood. One method that attempts to correct for this is a simple re-scaling of the model likelihood (Lohse et al., 2016). For example, making the assumption that every $n$th block is independent and dividing the likelihood by $n$ gives a more conservative estimate with which to compare likelihoods from a set of nested models. However, without information to guide the choice of $n$, choosing a value is arbitrary. The best method to correct for the affects of non-independence across blocks is to perform a parametric bootstrap (Lohse et al., 2016). When working with highly fragmented assemblies that consist of thousand of contigs, there is no accurate way of assessing the level of linkage between them. The bootstrap replicates correct for this by being simulated under a known level of linkage (e.g. here they are simulated as 20 continuous stretches of recombining sequence that, when combined, equal the total length of the contigs from the assembly). This allows the variance around the estimates to be assessed when accounting for this level of LD in the data.

The simulation program *msprime* (version 0.3.1 (IM) and 0.4.0 (ADM) (Kelleher et al., 2016)) was used to generate a hundred bootstrap datasets corresponding to the best-fit model under each scenario (IM and ADM). I used a python (Python 2) script (adapted from one written by K. Lohse and L. Bünnefeld) within which the scaled parameter estimates output by the models could be used as input specifications along with a recombination rate estimated from the data (see above) ($2.719 \times 10^{-10}$). The size and number of blocks matched those in the observed data and were simulated over 20 contiguous stretches of sequence per dataset. The simulated data were converted to the block format of four mutational types using the functions described in the block cutting section above. Each dataset was run through the applicable best-supported model in *Mathematica*. 95% confidence intervals (C.I) were estimated by taking 1.96 standard deviations (SD), either side of the MLEs.
3.3.1.7 Model selection – IM vs ADM

To assess whether a better fit to the data of one model over the other (in this case, the better fit of the ADM model) was significant, both the ADM and the IM model were fitted to 100 bootstrap datasets simulated under the best-fit IM model parameters. The differences in log likelihood (calculated as $\Delta l n L = 2*(ADMlnL - IMlnL)$) obtained from running the 100 replicates through the IM model and the ADM model were plotted as a histogram in *Mathematica*. A critical threshold significance value was obtained from this distribution at a significance level of 0.05 using the ‘Quantile’ function in *Mathematica*. If the $\Delta l n L$ obtained from the observed data was greater than the critical value, the ADM model was deemed to fit significantly better than the IM model.

3.3.2 PSMC analysis

3.3.2.1 Variant calling and consensus sequence generation

The PSMC infers the effective population size history of a single diploid individual. I ran one analysis for my two North haploid individuals and a second analysis for my two South haploid individuals. The authors of PSMC (Li and Durbin, 2011) recommend using *samtools* to call variants for use in the PSMC pipeline. To prepare the model input files, the per-individual recalibrated (BQSR) BAM files from the base recalibration part of the *GATK* pipeline (described in Chapter 2) were used as input for the *samtools mpileup* function (version 0.1.19). The resultant pileup file was converted to a VCF file in *bcftools* (version 0.1.19). A consensus sequence (fastq file) per individual was then generated using the *bcftools* utility *vcf2fq*. I used a pipeline suggested by the authors to create a diploid sequence for each of my individual pairs. First, the consensus files for each northern individual were checked to make sure they contained the same contig IDs and any that were not represented in both individuals were removed (this was repeated for the two southern individuals). This resulted in consensus files containing 31,371 and 33,523 contigs for the North and South pairs respectively. Next, the two consensus files per population were merged using the *seqtk* function *mergefa* (version 1.0) and converted to the PSMC input format using the PSMC function *fq2psmcfa* (version 0.6.5). The *fq2psmcfa* function is hard coded to include only contigs containing >10 kb of unfiltered bases and of these only those where 80% of the bases pass the quality threshold (in this case a base quality score of >20). This resulted in two input files ready to be run through the PSMC model (i.e. one northern ‘diploid’ individual and one southern ‘diploid’ individual). The two *psmcfa* input files contained 6900 and 7305 contigs for the North and South pairs respectively.
3.3.2.2 Running PSMC analyses

The input files were run through the PSMC program (version 0.6.5) using the following parameter values, N = 30, t = 20, r = 10.3 and p = “4+60*1+4”, where N is the number of iterations, t is the maximum coalescent timeframe for model inference (scaled by 2Ne), r is the initial $\theta/\rho$ ratio and p specifies how many time intervals the model spans. A range of different parameters were used to fit the model in an attempt to find the best trade-off between precision and over fitting (see Appendix D). Each dataset was sub-sampled 100 times to generate bootstrap replicates using the PSMC utility splitfa (version 0.6.5). The results were calibrated using the mutation rate stated above and a generation time of 4 per year. The results were plotted using a combination of the PSMC plotting utility psmc_plot.pl (version 0.6.5) and a custom python script that made use of the python modules pandas and matplotlib.
### 3.4 Results

#### 3.4.1 bSFS likelihood analyses

The block cutting process yielded 775,977 blocks of 387 bases, prior to four-gamete violation filtering. Removing the blocks that contained both Fixed and Shared sites (i.e. those that violated the 4-gamete test) resulted in a filtered dataset of 770,105 blocks (0.8% of the blocks were removed). Using the filtered data, the average per site heterozygosity for the North and South populations respectively was 0.000884 and 0.000822. Pairwise $F_{st}$ was 0.74, indicating high differentiation between the populations. Table 3.3 compares these properties across both the filtered and unfiltered datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. blocks</th>
<th>${A,B,F,S}$ proportions*</th>
<th>North/South heterozygosities</th>
<th>$F_{st}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfiltered</td>
<td>775,977</td>
<td>$[0.164,0.153,0.673,0.0102]$</td>
<td>$0.000953/0.000889$</td>
<td>0.73</td>
</tr>
<tr>
<td>Filtered</td>
<td>770,105</td>
<td>$[0.161,0.149,0.685,0.0050]$</td>
<td>$0.000884/0.000822$</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of the blockwise datasets pre and post filtering. *Proportions of $A$hets (North), $B$hets (South), Fixed and Shared sites.

#### 3.4.1.1 Isolation with continuous migration

The addition of post-divergence unidirectional migration results in a significantly better fit (i.e. the 95% confidence intervals (C.I.) around estimates of $M$ do not include zero) than models of strict divergence i.e. $M = 0$. Table 3.4 shows the log likelihood and parameter estimates for each model tested. Model support was strongest for migration occurring from the South into the North population in both the single $N_e$ and two $N_e$ models. Allowing for different $N_e$ values across the populations improves the IM model fit over the single $N_e$ scenario (the North population is inferred to be $\sim 4/5$ the size of the South population). Relative model support was assessed using likelihood ratio tests and AIC scores (Table 3.5). Both show that the model that includes gene flow and two different $N_e$ values fits the data significantly better than models without gene flow and/or with a single $N_e$. Figure 3.5 details the best supported model, for which 95% confidence intervals of maximum likelihood estimates (MLE) of parameters (obtained through the parametric bootstrap analysis) are detailed in Table 3.6. The migration rate is low ($M$...
Population genomics of a fig wasp community

= 0.071 (95% C.I – 0.045 – 0.097)), which equates to 1 migrant from South to North every 28 generations (equating to 1 wasp migrating every 7 years). The scaled parameter estimates and their confidence intervals are given in Table 3.6: the two populations diverged ~177 kya (95% C.I – 172 kya – 182 kya). The South population is inferred to have a larger Ne compared to the North (69,000 and 58,000 respectively).

<table>
<thead>
<tr>
<th>Model*</th>
<th>lnL</th>
<th>Diversity estimate parameter (θ)</th>
<th>Divergence time parameter (T)</th>
<th>Ne scaling parameter (B)</th>
<th>Migration rate parameter (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRDIV1Ne</td>
<td>-2.2545*10^6</td>
<td>0.337</td>
<td>3.938</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>IM1NeAB</td>
<td>-2.2367*10^6</td>
<td>0.287</td>
<td>5.210</td>
<td>n/a</td>
<td>0.044</td>
</tr>
<tr>
<td>IM1NeBA</td>
<td>-2.2328*10^6</td>
<td>0.278</td>
<td>5.533</td>
<td>n/a</td>
<td>0.057</td>
</tr>
<tr>
<td>STRDIV2Ne(b)</td>
<td>-2.2530*10^6</td>
<td>0.368</td>
<td>3.523</td>
<td>1.198</td>
<td>n/a</td>
</tr>
<tr>
<td>STRDIV2Ne(a)</td>
<td>-2.2544*10^6</td>
<td>0.346</td>
<td>3.808</td>
<td>1.055</td>
<td>n/a</td>
</tr>
<tr>
<td>IM2NeAB(b)</td>
<td>-2.2345*10^6</td>
<td>0.245</td>
<td>6.417</td>
<td>0.774</td>
<td>0.051</td>
</tr>
<tr>
<td>IM2NeBA(b)</td>
<td>-2.2322*10^6</td>
<td>0.255</td>
<td>6.193</td>
<td>0.872</td>
<td>0.061</td>
</tr>
<tr>
<td>IM2NeAB(a)</td>
<td>-2.2334*10^6</td>
<td>0.324</td>
<td>4.573</td>
<td>1.348</td>
<td>0.065</td>
</tr>
<tr>
<td>IM2NeBA(a)</td>
<td>-2.2317*10^6</td>
<td>0.299</td>
<td>5.137</td>
<td>1.193</td>
<td>0.071</td>
</tr>
</tbody>
</table>

Table 3.4: Log likelihoods and parameter estimates for each IM model tested. *AB signifies gene flow occurs from population A into population B, BA signifies gene flow occurs from population B into population A.
Population genomics of a fig wasp community

<table>
<thead>
<tr>
<th>Model</th>
<th>ΔlnL</th>
<th>LRT</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRDIV1Ne</td>
<td>-22,877.63</td>
<td>n/a</td>
<td>4.509*10^6</td>
</tr>
<tr>
<td>IM1NeAB</td>
<td>-5,078.23</td>
<td>35,598.8</td>
<td>4.473*10^6</td>
</tr>
<tr>
<td>IM1NeBA</td>
<td>-1,122.75</td>
<td>43,509.8</td>
<td>4.465*10^6</td>
</tr>
<tr>
<td>STRDIV2Ne(b)</td>
<td>-22,740.59</td>
<td>274.069</td>
<td>4.508*10^6</td>
</tr>
<tr>
<td>STRDIV2Ne(a)</td>
<td>-21,298.76</td>
<td>3,157.74</td>
<td>4.505*10^6</td>
</tr>
<tr>
<td>IM2NeAB(b)</td>
<td>-2,855.77</td>
<td>36,885.99</td>
<td>4.469*10^6</td>
</tr>
<tr>
<td>IM2NeBA(b)</td>
<td>-500.79</td>
<td>44,479.60</td>
<td>4.464*10^6</td>
</tr>
<tr>
<td>IM2NeAB(a)</td>
<td>-1,757.83</td>
<td>39,081.86</td>
<td>4.466*10^6</td>
</tr>
<tr>
<td>IM2NeBA(a)</td>
<td>0</td>
<td>45,481.19</td>
<td>4.463*10^6</td>
</tr>
</tbody>
</table>

Table 3.5: Relative model support for the IM models. Difference in log likelihood (ΔlnL) calculated as the difference between model lnL and best fit model lnL. Likelihood ratio tests (LRT) calculated as, e.g. LRT = 2*(IM2NeBAxlnL-STRDIV2NexlnL). AIC scores calculated as - 2lnL + 2*(number of independently adjusted parameters within the model).
Figure 3.5: Diagrams of the best supported IM and ADM models. Each are scaled according to the scaled divergence time (and in the case of the ADM model the admixture time) estimates in years. The 95% C.I for the divergence times are indicated by the thin black lines. The populations are scaled according to their relative difference in $N_e$ estimates.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.299 (0.283-0.314)</td>
<td>0.300 (0.286-0.315)</td>
<td>69.029 (65,414-72,643)</td>
<td>69.321 (66,059-72,583)</td>
<td>0.299</td>
<td>0.302</td>
<td>0.299</td>
<td>0.300</td>
<td>69.029</td>
<td>69.321</td>
<td>0.299</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>$B$</td>
<td>1.193 (1.089-1.296)</td>
<td>1.181 (1.093-1.269)</td>
<td>57,871 (54,285-61,457)</td>
<td>58,690 (55,212-62,167)</td>
<td>1.197</td>
<td>1.184</td>
<td>1.197</td>
<td>1.184</td>
<td>57,871</td>
<td>58,690</td>
<td>1.197</td>
<td>1.184</td>
</tr>
<tr>
<td></td>
<td>$T^*$</td>
<td>5.137 (4.784-5.489)</td>
<td>5.643 (5.317-5.966)</td>
<td>709,135 (689,995-728,275)</td>
<td>782,103 (771,061-793,146)</td>
<td>5.150</td>
<td>5.604</td>
<td>5.150</td>
<td>5.604</td>
<td>709,135</td>
<td>782,103</td>
<td>5.150</td>
<td>5.604</td>
</tr>
<tr>
<td></td>
<td>$T^{^*}$</td>
<td>N/A</td>
<td>N/A</td>
<td>177,284 (172,499-182,069)</td>
<td>195,526 (192,765-198,286)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>177,284</td>
<td>195,526</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>$T_1$</td>
<td>N/A</td>
<td>1.656 (1.509-1.802)</td>
<td>N/A</td>
<td>229,554 (212,752-246,356)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>229,554</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>$T_1^{^*}$</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>57,388 (53,188-61,589)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>57,388</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>$M^{^*}$</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>2.573<em>10^-7 (1.639</em>10^-7-3.506*10^-7)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>2.573*10^-7</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>$f^{^\text{^*}}$</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.239</td>
</tr>
</tbody>
</table>

Table 3.6: Details of the parameter MLEs for the best supported IM and ADM models with their corresponding 95% C.I in brackets (top two rows). The scaled parameter MLEs with their corresponding 95% C.I in brackets (middle two rows). The mean of each parameter estimate across bootstrap replicates (bottom two rows). *estimates of divergence/admixture times in generations. **estimates of divergence/admixture times in years. ^^migration rate. ^admixture fraction.
3.4.1.2 Isolation with instantaneous admixture

The addition of a post-divergence unidirectional admixture event to both the single $Ne$ and two $Ne$ models produces significantly improved fits over strict divergence models. Table 3.7 shows the log likelihood and MLEs of parameters for each model. The best-supported direction for admixture is from the South into the North population. This is true of both the single and two $Ne$ models. The addition of the extra $Ne$ parameter improved model fit over the equivalent scenario inferring a single $Ne$ for both populations (the North population is again inferred to be $\sim 4/5$ the size of the South population). The admixture fraction inferred from the best-fit model suggests around a quarter ($f = 0.239$) of North lineages trace back to the South population during this event. Both likelihood ratio tests and AIC scores (Table 3.8) show a significantly better fit for the models that include the admixture event and two different $Ne$ values compared to the models without admixture and with a single $Ne$ value, respectively. Figure 3.5 details the best-supported ADM model, the 95% confidence intervals for MLEs of each parameter are detailed in Table 3.6. The scaled parameter estimates and their confidence intervals are given in Table 3.6. The model infers that the two populations diverged $\sim 196$ kya (95% C.I – 193 kya – 198 kya). The South population is inferred to have a slightly larger $Ne$ compared to the North (69,000 and 59,000 respectively). The admixture event is inferred to have occurred, South to North, $\sim 57$ kya (95% C.I – 53 kya – 62 kya).
Population genomics of a fig wasp community

<table>
<thead>
<tr>
<th>Model*</th>
<th>lnL</th>
<th>Diversity estimate parameter (θ)</th>
<th>Admixture time parameter (T1)</th>
<th>Divergence time parameter (T2)</th>
<th>Ne scaling parameter (B)</th>
<th>Admixture fraction (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADM1NeT10AB</td>
<td>-2.2359*10^6</td>
<td>0.287</td>
<td>n/a</td>
<td>4.969</td>
<td>n/a</td>
<td>0.015</td>
</tr>
<tr>
<td>ADM1NeT10BA</td>
<td>-2.2359*10^6</td>
<td>0.287</td>
<td>n/a</td>
<td>4.969</td>
<td>n/a</td>
<td>0.015</td>
</tr>
<tr>
<td>ADM1NeAB</td>
<td>-2.2346*10^6</td>
<td>0.286</td>
<td>2.194</td>
<td>3.928</td>
<td>n/a</td>
<td>0.296</td>
</tr>
<tr>
<td>ADM1NeBA</td>
<td>-2.2307*10^6</td>
<td>0.279</td>
<td>1.970</td>
<td>4.215</td>
<td>n/a</td>
<td>0.257</td>
</tr>
<tr>
<td>ADM2NeT10AB</td>
<td>-2.2361*10^6</td>
<td>0.330</td>
<td>n/a</td>
<td>4.193</td>
<td>1.313</td>
<td>0.014</td>
</tr>
<tr>
<td>ADM2NeT10BA</td>
<td>-2.2352*10^6</td>
<td>0.306</td>
<td>n/a</td>
<td>4.631</td>
<td>1.152</td>
<td>0.017</td>
</tr>
<tr>
<td>ADM2NeAB(b)</td>
<td>-2.2312*10^6</td>
<td>0.323</td>
<td>1.642</td>
<td>3.553</td>
<td>1.329</td>
<td>0.257</td>
</tr>
<tr>
<td>ADM2NeBA(b)</td>
<td>-2.2297*10^6</td>
<td>0.256</td>
<td>2.003</td>
<td>4.869</td>
<td>0.857</td>
<td>0.256</td>
</tr>
<tr>
<td>ADM2NeAB(a)</td>
<td>-2.2315*10^6</td>
<td>0.247</td>
<td>2.281</td>
<td>4.995</td>
<td>0.767</td>
<td>0.291</td>
</tr>
<tr>
<td>ADM2NeBA(a)</td>
<td>-2.2296*10^6</td>
<td>0.300</td>
<td>1.656</td>
<td>3.985</td>
<td>1.181</td>
<td>0.239</td>
</tr>
</tbody>
</table>

Table 3.7: Log likelihoods and parameter estimates for each ADM model tested. *AB signifies gene flow occurs from population A into population B, BA signifies gene flow occurs from population B into population A.

<table>
<thead>
<tr>
<th>Model</th>
<th>ΔlnL</th>
<th>LRT</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADM1NeT10AB</td>
<td>-6,349.88</td>
<td>37,230.1</td>
<td>4.472*10^6</td>
</tr>
<tr>
<td>ADM1NeT10BA</td>
<td>-6,349.88</td>
<td>37,230.1</td>
<td>4.472*10^6</td>
</tr>
<tr>
<td>ADM1NeAB</td>
<td>-4,994.89</td>
<td>2,633.9</td>
<td>4.469*10^6</td>
</tr>
<tr>
<td>ADM1NeBA</td>
<td>-1,110.12</td>
<td>10,478.2</td>
<td>4.461*10^6</td>
</tr>
<tr>
<td>ADM2NeT10AB</td>
<td>-6,536.46</td>
<td>36,582.8</td>
<td>4.472*10^6</td>
</tr>
<tr>
<td>ADM2NeT10BA</td>
<td>-5,631.81</td>
<td>35,508.5</td>
<td>4.470*10^6</td>
</tr>
<tr>
<td>ADM2NeAB(b)</td>
<td>-1,644.65</td>
<td>9,783.6</td>
<td>4.462*10^6</td>
</tr>
<tr>
<td>ADM2NeBA(b)</td>
<td>-172.39</td>
<td>10,918.84</td>
<td>4.460*10^6</td>
</tr>
<tr>
<td>ADM2NeAB(a)</td>
<td>-1975.22</td>
<td>9,122.49</td>
<td>4.463*10^6</td>
</tr>
<tr>
<td>ADM2NeBA(a)</td>
<td>0</td>
<td>11,263.6</td>
<td>4.459*10^6</td>
</tr>
</tbody>
</table>

Table 3.8: Relative model support for the ADM models. Difference in log likelihood (ΔlnL) calculated as the difference between model lnL and best fit model lnL. Likelihood ratio tests (LRT) calculated as, e.g. LRT = 2*(ADM2NeBAxlnL - ADM2NeT10lnL). AIC scores calculated as -2lnL + 2*(number of independently adjusted parameters within the model).
3.4.1.3 Assessing model fit

To assess the relative fit of each model to the data, the expected frequencies of Ahets, Bhets, Fixed and Shared sites under the best fitting IM and ADM scenarios were compared to the observed frequencies (Figure 3.6). The expected frequencies under both models differ only slightly, over each mutation category, from the observed frequencies. The data show a higher than expected frequency of Ahets, Bhets and Fixed differences, under both models. The expected frequencies of Shared hets are reduced under both models compared to that observed in the data. The expected number of Shared hets under the IM model (2410) is roughly a third of that observed in the data (7937) with the expected number under the ADM model (878) being roughly a third of that expected under the IM model. This suggests that the IM model fits the data better than the ADM model in terms of the Shared heterozygous sites, however, the 95% confidence intervals around the expected frequencies under both models overlap signifying the difference between these two frequencies is not significant. The confidence intervals around both the expected frequencies of Shared heterozygous sites do not overlap with the observed frequency suggesting the frequency of Shared hets observed in the data is significantly more than expected under either model. The expected bSFS (i.e. the expected table of configuration counts) was also inferred under each best fit model and plotted against the observed bSFS to help assess model fit (Figure 3.7). The fits for both models are reasonable, however a few configurations stand out as deviating from expectations (i.e. deviating from the diagonal in Figure 3.7). For both models the configurations that deviate from a ‘perfect’ fit fall on the same sides of the diagonal. In particular, the two most common configurations {0,0,1,0} (i.e. one fixed difference) and {0,0,0,0} (i.e. monomorphic blocks) in the observed data are encountered more often than expected under both the IM and ADM models.
Figure 3.6: The expected frequencies of Ahets, Bhets, Fixed and Shared sites under the best fitting IM and ADM models alongside the observed frequencies.
Figure 3.7: The expected bSFS under the best fit IM (top) and ADM (bottom) models plotted against the observed bSFS. The grey diagonal line indicates what would constitute a perfect fit.

3.4.1.4 Model selection – IM vs ADM

The ADM model has a higher log likelihood compared to the IM model ($\Delta \ln L = 4175$). This might simply reflect the fact that the ADM model is more complex, i.e. contains an additional parameter. As the models are not nested, it is not possible to assume a chi-squared distribution and use a standard likelihood ratio test for the difference in lnL ($\Delta \ln L$). Instead, simulations generated under the IM model
parameters were used to obtain a distribution (and from this a critical value) of the difference in log likelihoods between the IM and ADM models (Figure 3.8). This approach asks whether the simpler history can be rejected in favour of a history of discrete admixture. The distribution generated here (Figure 3.8) centres around zero with a critical value of 155.4 (p=0.05). The distribution being centred around zero suggests that if the true history was one of continuous gene flow (IM) there would not be enough signal in the data to distinguish between the two modes of gene flow. However, the fact that the ΔlnL observed from the data is much greater than the critical value supports the conclusion that the ADM model does fit the data significantly better than the IM model.

Figure 3.8: Histogram showing the distribution of the difference in log likelihoods between the best fit IM and ADM models.

3.4.2 PSMC analyses

The PSMC plots were scaled using the *D. melanogaster* mutation rate (2.8*10^-9) used above and a generation time of 0.25 years (4 per year, as assumed above). Figure 3.9(a) shows the inferred Ne trajectories for both populations looking back in time. Overlaid on Figure 3.9(c) are the inferred Ne estimates for each population along with the divergence time and the time of the admixture event from the best-supported ADM model. Overlaid on Figure 3.9(b) are the inferred Ne estimates for
each population along with the divergence time from the best-supported IM model. The inferred divergence times from both models occur where the two trajectories have very similar Ne values (the bootstrap replicates overlap). However, prior to the inferred divergence times (going back in time) and during the most recent time interval (~1,000-5,000 ya) the bootstrap replicates widen drastically around the inferred trajectory. This suggests the method is unable to infer the recent and more ancient events due to a lack of coalescent events occurring within the samples. This is confirmed by the authors of PSMC, who state this will always be a problem when inference is made using small samples (Li and Durbin, 2011). Inference of recent events is also restricted by assembly quality. Fragmented assemblies contain large numbers of short contigs, many of which will not have been included in the model due to filtering cut-offs (78% of contigs were lost from both pairs due to the filtering criteria), limiting inference to the remaining blocks of sequence of suitable length. The Ne estimates inferred by both the IM and ADM model agree closely with the PSMC inferred values around the time of divergence, with the North being slightly smaller than the South population. Just prior to the time of the inferred admixture event (Figure 3.9(c)) the two Ne trajectories start to separate with an increase in the northern Ne and a decrease in the southern. This change could be accounted for by the presence of gene exchange from the southern population into the north inferred by the ADM model. An influx of migrants from one population to another would increase the levels of genetic diversity present in the receiving population and in turn increase the populations Ne.
Population genomics of a fig wasp community

(a) $P. nigriventris$ PS MC plot

(b) $P. nigriventris$ PS MC plot - IM

(c) $P. nigriventris$ PS MC plot - ADM

Years ($g = 0.25, \mu = 2.8 \times 10^{-9}$)
Figure 3.9: The $Ne$ trajectories inferred from the PSMC analysis including the bootstrap replicates (a). Overlaid on the plot (b) are the North and South $Nes$ (inclusive of their 95% C.I - horizontal grey lines) and the divergence time (inclusive of its 95% C.I - vertical grey lines) inferred from the IM model. Overlaid on the plot (c) are the North and South $Nes$ (inclusive of their 95% C.I - horizontal grey lines) and the divergence and admixture times (inclusive of their 95% C.I – vertical grey lines) inferred from the ADM model. The model is calibrated using a mutation rate ($\mu$) of $2.8 \times 10^{-9}$ and a generation time of 0.25 (4 generations per year).
3.5 Discussion

3.5.1 Population history of *P. nigriventris*

This study uses low coverage genome-wide data to infer the divergence history between two populations of the rainforest dwelling pollinating fig wasp, *P. nigriventris*. The analyses presented here support the hypotheses laid out in the objectives, inferring significant divergence between the two populations in the late Pleistocene. Although significant support was established for the ADM model over the IM model it is important to note that both models agree closely in values for the key parameters they infer. Both infer the presence of significant post-divergence gene flow between the two populations and agree in showing the strongest support for gene flow from the South population into the North. They also agree by supporting different effective population sizes, larger in the South than in the North. This overall history is reinforced by the two *N*\textsubscript{e} trajectories inferred from the PSMC analysis. The late Pleistocene divergence inferred under both the IM and ADM models occurs when the bootstrap replicates for the two trajectories overlap (Figure 3.9). Prior to divergence the PSMC bootstrap replicates show a very large variance in *N*\textsubscript{e} suggesting there are not enough coalescent events for accurate inference. Post divergence, the North *N*\textsubscript{e} increases whilst the South goes through a period of decline. This is consistent with gene flow being predominant from the South into the North population. The subsequent decline in the North population *N*\textsubscript{e} however suggests the period of gene flow into the North was temporary. This pattern could also be indicative of a more severe contraction of habitat area in the South (McGuigan et al., 1998) during this period, reducing the range (and hence size) of the South population whilst in the North a larger range was maintained.

The IM model infers a constant rate of post-divergence gene flow from the South into the North population. The inferred migration rate (*M* = 0.071) equates to 0.036 migrants per generation (0.14 migrants per year). The ADM model infers an instantaneous admixture event at ~57 kya in which nearly a quarter of the lineages in the South moved into the North population. This admixture event could be indicative of a time when a change in conditions allowed the temporary dispersal of individuals between isolated populations. Having established support for significant gene flow, it would be interesting to be able to assess whether enough signal was present to fit divergence models with differing gene flow scenarios (e.g. continuous gene flow for a limited period post-divergence (Wilkinson-Herbots, 2012; Costa and Wilkinson-Herbots, 2017)) to test this hypothesis.
The concordant gene flow direction inferred by both models does not rule out bidirectional dispersal but rather suggests a signal of predominant gene flow from the South into the North. The reason for this is unclear but could be affected by dispersal conditions being more favourable in this direction (e.g. wind direction) and/or source sink dynamics (e.g. if the South harbours a larger geographic population, individuals are more likely to disperse into the North) (Oswald et al., 2017). Current (1940s-2000s) wind directions suggest an overall trend towards stronger winds predominantly coming from the South (Australian Government, Bureau of Meteorology, http://www.bom.gov.au/climate/) however, this varies with season and time of day and, more importantly, there are no details of these sorts of weather patterns over the longer term historical timescales considered here. The larger Ne inferred for the South population could be indicative of a larger geographic population however the difference between the two Nes is small (North is ~4/5 the size of the South) and the tree density is thought to currently be similar in both areas (Dixon, 2003, J. Cook, personal communication). On top of this, there are a number of factors that effect the Ne of a population (e.g. unequal sex ratios and variation in offspring numbers (Nielsen and Slatkin, 2013)), it does not necessarily correlate with the actual population size.

3.5.2 Have the Burdekin and St. Lawrence Gaps contributed to the divergence history of *P. nigriventris*?

This phylogeographic study spans both the Burdekin and St. Lawrence gaps on the east Queensland coast of Australia. The mesic habitat found along the east coast is the most extensively studied area of the continent (Edwards and Melville, 2010), however few studies have spanned the region of this study (particular focus has been on the Wet Tropics in north-east Queensland (e.g. Schneider et al., 1998; Moritz et al., 2000; Pope et al., 2000; Bell et al., 2007; Krosch et al., 2009; Moritz et al., 2009; Rossetto et al., 2009; Mellick et al., 2014; Moreau et al., 2015)) and even fewer have used genomic scale data to assess phylogeographic relationships. Despite Australia not experiencing much direct glaciation during the climatic oscillations of the Plio-Pleistocene, it has a complex climate history, a major feature of which is the transition towards aridity that started in the Miocene and continued throughout the Plio- and Pleistocene (Schäuble and Moritz, 2001; Martin, 2006; MacQueen et al., 2010; Frankham et al., 2016). These periods of long term drying, characterised by warmer, wetter interglacials and cooler, drier glacial, are thought to have caused habitat contractions in rainforest taxa and expansions in dry forest adapted taxa (Byrne, 2008; Maldonado et al., 2012; Burke et al., 2013). Evidence from pollen,
fossils and paleoclimate models suggest the glacial cycles of the Pleistocene had a particularly marked effect on the rainforests along the east coast, which would have been more prevalent pre-Pleistocene (McGuigan et al., 1998; Schneider et al., 1998; Pope et al., 2000; Bell et al., 2007). The dry habitat corridors known as the Burdekin and St. Lawrence gaps are assumed to be relics of this long term drying. However, there is no strong evidence to confirm when they were formed (Bryant and Krosch, 2016). The effects of these biogeographic barriers have been investigated in a range of taxa, including mammals, plants, reptiles and invertebrates (Schiffer et al., 2007; Edwards and Melville, 2010; Lucky, 2011; Burke et al., 2013; Smissen et al., 2013; Bryant and Fuller, 2014). The presence of divergence across the gaps is not uniformly supported across these studies and the associated divergence times (where estimated) range from early Miocene through to late Pleistocene (Dolman and Moritz, 2006; Baker et al., 2008; Edwards and Melville, 2010; Catullo and Scott Keogh, 2014). Furthermore, the majority of these divergence time estimates have been inferred from single gene (primarily mtDNA) phylogenies, which can severely overestimate population divergence. This is because they represent the time to the most recent common ancestor of that particular gene (or gene segment), which has to predate population divergence. If the populations were large in the past, such overestimates of divergence age could be substantial. The discordance in the inferred effects of these gaps on divergence events is therefore not necessarily surprising given the range of taxa and the differences in their associated ecologies and effective population sizes. However, it is difficult to compare divergence times across studies that use different markers, methods and sampling schemes to estimate vicariance. A Pleistocene divergence, as inferred here, has also been proposed in *Melomys cervinipes* (a wet forest restricted rodent) and *Varanus varius* (a large lizard with broad habitat preferences) (Smissen et al., 2013; Bryant and Fuller, 2014). Each of these studies used both a small number of sequence loci (two mtDNA (16S and cyt b) and one nuclear (AP5), and one mtDNA (ND4), respectively) to reconstruct phylogenies, and eight microsatellites to assess genetic structure, across the sampled individuals. Bryant and Fuller (2014) used calibration points from a previously inferred phylogeny to place divergence time estimates on the major clades in their three gene fragment phylogeny. Smissen et al (2013), however, used three fossil calibrations to estimate the divergence times for each clade in their single mtDNA phylogeny. Studies of invertebrates along the east coast have covered a range of taxa (e.g. Arachnida (Rix and Harvey, 2012), Insecta (Eastwood et al., 2006; Bell et al., 2007) and Oligochaeta (Moreau et al., 2015)) but few have estimated divergence across the Burdekin and St. Lawrence gaps in insects (Baker et
al., 2008). Schiffer et al (2007) assessed phylogeographic patterns of *Drosophila birchii* from the tropical rainforests of Queensland using ten microsatellite loci and a single mtDNA locus (ND5). They found no evidence for genetic structure across the Burdekin gap and suggest *D. birchii*, a rainforest habitat specialist, may have undergone a recent range expansion following periods of glacial contraction. They also inferred a moderate level of gene flow across the distribution of *D. birchii* and propose this could be facilitated by long distance dispersal via wind currents and the ability to survive in habitats outside their rainforest distribution. These results are in contrast to those in this study for *P. nigriventris*, which is also a rainforest specialist, but one expected to have a much higher affinity for its rainforest host given the obligate mutualism that exists between them. It is impossible to infer the location of the ancestral population of *P. nigriventris* from the analyses presented here: it could have originated in the North or the South or covered both regions prior to the divergence event. As the existence of more extensive rainforest cover along the east coast has been inferred for Australia’s pre-Pleistocene history (Martin, 2006), for the sake of the following speculations I will assume the ancestral population was widely spread, covering both of the present day population areas. The ADM population history inferred here for *P. nigriventris* is consistent with a scenario in which divergence began following a particularly harsh contraction of suitable habitat during the severe climatic cycles of the Pleistocene, and was followed by an extended period during which gene flow was almost non-existent. Then, during a period of more favourable conditions, the distribution of these rainforest trees expanded, allowing exchange of migrants before subsequent contractions returned their distribution to what it is today. It is possible that the pre-admixture distributions were more extensive than they are now, increasing the likelihood of gene flow during more favourable conditions, and that post-admixture the populations have contracted further due to increasing aridity and/or other influences (e.g. the arrival of humans, landscape burning).

Hocknull et al (2007) analysed the first suite of vertebrate fossils from eastern Queensland dating from the middle Pleistocene. They present evidence of a major faunal turnover in rainforest taxa between ~280 - 205 kya, a period characterised by decreased precipitation and more severe aridification than in earlier periods. This timing supports the hypothesis that such climatic changes could have influenced the divergence of *P. nigriventris*. Despite the date of the inferred admixture pulse (~57 kya (95% CI – 53 kya – 62 kya)) for *P. nigriventris* being in the middle of the last 100,000 year glacial cycle, marine isotope analyses recognise MIS3 (Marine Isotope
Stage 3 (27 kya – 60 kya) as a period characterised by abrupt warming phases (Siddall et al., 2008; Van Meerbeeck et al., 2009). These warming phases are interspersed by cooler periods, the lowest temperatures of which are not thought to drop to those of the Last Glacial Maximum (LGM – ~19 kya – 21 kya) (Van Meerbeeck et al., 2009). Clark et al (2007) suggest these MIS3 climate variations occur in ~7,000 year cycles, the lower bound (~53 kya) of the admixture time estimated here roughly corresponding to the warmest point in one of these cycles. These abrupt climate change phases occurred during a period of major flora and faunal change on the Australian continent, which included the arrival of humans, an increase in landscape burning, continued movement towards aridity and the extinction of Australia’s megafauna (Hocknull et al., 2007; Wroe et al., 2013). It is not exactly clear how these different historical events have influenced the distribution of *F. watkinsiana* and its pollinator *P. nigriventris*, but the results presented here make it seem highly likely that its history has been shaped by the extreme habitat changes that occurred during the mid and late Pleistocene.

**3.5.3 Limitations of the demographic inference methods**

**3.5.3.1 Violations of the bSFS scheme assumptions**

Multi locus inference methods, including the likelihood calculation based on the bSFS used here, assume no intralocus (i.e. within-block) recombination and free interlocus recombination (i.e. between blocks). These assumptions could bias and result in overconfidence in the parameter estimates inferred, respectively. The presence of intralocus recombination could result in confounding genealogical histories occurring within a block that will decrease observed coalescent variance and could lead to a bias in the parameter estimates (Hare, 2001; Felsenstein, 2003; Wall, 2003). For example, it has been shown that it could lead to decreased estimates of $N_e$, increased estimates of divergence time and decreased estimates of migration (Wall, 2003; Strasburg and Rieseberg, 2010; Nurnberger et al., 2016). In contrast, however, several studies have shown parameter estimates to be robust to violations of the assumption (Jennings and Edwards, 2005; Lanier and Knowles, 2012; Hearn et al., 2014; Bunnefeld et al., 2015; Wang and Liu, 2016). The overconfidence associated with violations of the free interlocus recombination assumption stems from the likelihood of each model being calculated across blocks that are assumed to be independent, which when untrue essentially results in a composite likelihood calculation (Lohse et al., 2016). The best way to account for violations to these assumptions, given the approach I have used, is to perform a parametric bootstrap. Such analyses are not common in demographic inference studies due to their
computationally demanding nature. The bootstrapping procedure was carried out for both the best fit IM and ADM models. The 95% confidence intervals are narrow for each parameter under both models and the mean estimates obtained from the simulation replicates are very close to the MLEs used to simulate the datasets, suggesting little bias in the estimates (Table 3.6). Also, support from the likelihood ratio tests, performed to assess the fit of the series of nested models (Tables 3.5 and 3.8), remains significant even when the overinflated log likelihoods are re-scaled assuming that only every 100th or every 1000th block is truly independent (i.e. dividing the log likelihoods by 100 and 1000 respectively and assuming a chi-squared distribution with the appropriate degrees of freedom). To further explore the effects on parameter estimates of violating these assumptions, data could be simulated under a range of recombination rates and partitioned using a range of block lengths.

A key assumption of the bSFS scheme is that the effective neutral mutation rate ($\mu$) is equal across blocks. The data generated here were sampled with the intention of generating a random genome-wide sample per individual. However, the assembly is likely to contain biases due to the complex nature of the genome (e.g. coding vs non-coding, repetitive elements). Coding regions are expected to be constrained by selection and therefore are likely to show less variation across individuals. A consequence of this is that these regions are likely to assemble and align preferentially, potentially resulting in a dataset that is enriched for more conserved coding regions. The signature of a species’ demographic history is common to all neutral loci whereas the effect of selection is localised (Hare, 2001). To assess whether the estimates were being biased by the presence of selective constraint, the genomic regions likely under selection, i.e. coding regions, could be removed from the data and the analysis re-run. Hearn et al (2014) attempted to account for the violation of the equal $\mu$ across blocks assumption by partitioning loci by their respective proportion of coding sequence and scaling the estimated genome-wide mutation rate accordingly. Encouragingly they found that not accounting for mutation rate heterogeneity did not change the best fit models inferred and had little affect on the parameters estimated, reporting only a slight increase in Ne and divergence time estimates. However, assessing the proportion of coding to non-coding regions is difficult in species that are poorly covered in sequence databases and for which there is no annotated genome or transcriptome to aid in gene detection (as is the case here). One potential indication of biased data sampling is the presence of a higher than expected frequency of monomorphic blocks. The
Populations genomics of a fig wasp community

 Frequencies of invariant blocks expected under the neutral best fit IM and ADM models were calculated and compared to the observed frequency in the data. This demonstrated an excess of monomorphic blocks in the data under both scenarios. Balancing selection is expected to have the same effect on a coalescent tree as a history of population sub-division with low levels of gene flow (Hein et al., 2005). Both situations are expected to maintain a level of shared polymorphisms, but, in the case of sub-division these sites should be distributed across the sampled loci (unless gene flow is very recent), whereas, in the case of balancing selection these sites should be restricted to only a few loci. To assess this in the *P. nigriventris* dataset, the distribution of shared heterozygous sites across all blocks (filtered for 4-gamete test violations) was plotted (Figure 3.10). This shows that, out of the blocks that contain shared het sites, most blocks contain small numbers (87% of blocks contain 1 or 2 shared het sites) and there is no peak around blocks with large numbers of shared heterozygous sites. Also, the observed distribution of shared heterozygous sites shows a slight increase in shared het sites compared to the analytic expectation under both the IM and ADM models (Figure 3.10). Divergent selection on the other hand could lead to more fixed differences between the population than would be expected under drift resulting in an overestimate of the divergence time. As with balancing selection, divergent selection would effect specific loci whereas an old population split would affect divergence across the whole genome. To assess this in the *P. nigriventris* dataset, the same process as above was carried out but for the divergent sites (i.e. the fixed differences) (Figure 3.11). The distribution of divergent sites observed in the data shows that the majority of blocks contain few fixed differences and there is no excess of blocks with large numbers of fixed differences. The expected distributions under both the IM and ADM histories show that both models give a tight fit to the data.
Figure 3.10: The observed distribution of Shared heterozygous sites across all blocks (solid black line). The expected distribution of Shared heterozygous sites under the IM (dashed line) and ADM (dotted line) models.

Figure 3.11: The observed distribution of divergent sites (i.e. Fixed differences) across all blocks (solid black line). The expected distribution of divergent sites under the IM (dashed line) and ADM (dotted line) models.

3.5.3.2 Are PSMC analyses robust to parameters used?
PSMC was developed primarily for use with human data and so the recommended input parameters (-N 25, -t 15, -r 5 and -p “4+25*2+4+6”) are only appropriate for use with data expected to have similar properties (Li and Durbin, 2011). Without prior knowledge regarding parameter choice (with the exception of the $\theta / \rho$ ratio (-
r) that was calculated using the estimates from the bSFS analysis), a range of parameters (Appendix D) were tested based on those used previously in other studies (e.g. Groenen et al., 2012; Wallberg et al., 2014; Meyer et al., 2015; Chen et al., 2016; Nadachowska-Brzyska et al., 2016; Small et al., 2016) and the expectation that the signal would diminish prior to the divergence time inferred from the bSFS models. While the level of precision varied, the overall patterns of the two Ne trajectories were robust across the range of parameters tested (Appendix D).

### 3.5.3.3 Generation time

Both the bSFS and PSMC analyses were calibrated using the same generation time. The generation time for *P. nigriventris* is not known, in fact, estimates of generation times in fig wasps in general are scarce. Liu et al (2014a) estimate 2-3 generations in a year for a pollinator of a dioecious fig tree in eastern China but I am unaware of any in monoecious species. The shortage in estimates is primarily due to their dependence on their fig tree hosts which fruit asynchronously year round (Harrison, 2005), meaning there is no clear seasonal cycle to aid generation time discovery in contrast to, for example, cynipid gall wasps (Stone et al., 2002). Through personal observations, James Cook and Tim Sutton (Western Sydney University) estimate a likely range of 2-6 generations per year, with 4 being most likely. Given that there is more than 1 generation per year, the divergence time in generations puts an upper bound on the scaled time in years. Despite this uncertainty, the overall conclusion of this study that divergence occurred in the late Pleistocene for *P. nigriventris* holds true no matter which of these generation times is chosen.

### 3.5.4 Conclusions

This study demonstrates our ability to infer the population history of a previously unstudied naturally occurring pollinating fig wasp using low coverage, fragmented genome-wide data. It highlights that *P. nigriventris* split into two discrete populations in the late Pleistocene, followed by at least one pulse of gene flow. Divergence occurred during a period of particularly fluctuating climatic conditions in eastern Australia that likely affected the distribution of *F. watkinsiana* and its pollinator, *P. nigriventris*. 


4 Chapter 4: Comparative phylogeography of the inhabitants from a geographically widespread fig wasp community

4.1 Abstract

Comparative phylogeography is an important approach for investigating shared responses in co-distributed taxa. Historically, it has been based on data from single loci, predominantly mtDNA in animals, however, in the genomic era there is the potential to make comparisons across species using hundreds or thousands of loci. The methods available for making species comparisons under a unified statistical framework are limited, meaning that a researcher often has to choose between a joint analysis that includes all sampled species but sacrifices detail at the individual species level, or inferring detailed population histories per species and comparing them outside a specific statistical framework. The communities of wasps associated with fig trees make excellent model systems for community level studies due to their high host specificity and the fact that they all reproduce within the confines of the fig fruits. I assembled datasets for eight wasp species associated with two Australian fig trees, *F. rubiginosa* and *F. obliqua*, with the aim of comparing population histories within and across trophic levels along the east coast of Australia. Low coverage genome-wide data was generated for 4-6 individuals per species and used to infer their population histories under a composite likelihood scheme. The individual histories were then interpreted with respect to potential climatic drivers and compared within and across trophic levels to assess the level of concordance/discordance in this fig wasp community. The results show 7 out of the 8 species diverged in the Pleistocene, with the eighth showing no significant divergence between its populations. Within the Pleistocene however, the divergence times inferred varied widely, spanning from 24 kya to 1.4 mya. Although the patterns seen across the community do not fit with one single model of community assembly, several interesting patterns emerge from the data. The pollinator of *F. rubiginosa* shows no evidence of ongoing gene flow whereas both its parasitoid and inquiline show significant gene flow indicating they could be more effective dispersers. The hyperparasitoid shows a younger divergence time compared to its host parasitoid suggesting it may have followed it through time. A within guild comparison between the two inquilines from different host trees show strikingly similar histories, possibly reflecting a shared response to climatic conditions within
Population genomics of a fig wasp community
	his trophic level. Overall these results offer a tantalising glimpse into the history of this diverse community.
4.2 Introduction

4.2.1 Comparative phylogeography

Comparative phylogeography aims to assess the extent to which co-distributed taxa show similar population histories over space and time and, where possible, to link such patterns to shared exposure to dispersal barriers or climate and habitat fluctuations (Bermingham and Moritz, 1998; Hickerson et al., 2010). Comparative phylogeography is increasingly seen as an important tool for asking questions about how natural communities assemble (Hickerson et al., 2010; Smith et al., 2011). Over space and time, do members of a community share responses to environmental fluctuations? Or do subsets (such as different taxa, or trophic levels) respond differently? The answers to such questions are crucial to understanding current patterns of biodiversity and to predicting how future changes in climate or habitat structure may affect biological communities.

Historically, phylogeographic relationships amongst sets of interacting taxa have been compared using gene genealogies for one gene (predominantly mtDNA loci in animals, more often nuclear ITS or chloroplast regions in plants) (Edwards et al., 2015; Avise et al., 2016) or a small number of genes or other markers (such as microsatellites) to directly infer the presence of concordant or discordant genetic structure (Avise et al., 1987; Bermingham and Moritz, 1998; Hickerson et al., 2010). Single locus methods were soon realised to be problematic however as they do not take into account the stochastic variance in gene trees due to genetic drift (Knowles, 2009). This coalescent variance presents two kinds of problems for a single gene approach. First, populations differing in effective population size and/or mutation rate can show very different gene trees despite sharing important aspects of their history, such as population divergence times (Knowles, 2009). Importantly however, the stochastic nature of the coalescent means gene trees can also differ between species just by chance even if mutation rates and effective population sizes ($N_e$) are shared across taxa (Rosenberg and Nordborg, 2002). Second, species that have different underlying histories can show misleadingly similar gene tree structures (Edwards and Beerli, 2000; Hey and Machado, 2003). A particular problem is that divergence times (i.e. node ages) in a gene tree are influenced in confounding ways by $N_e$ and the actual splitting time between populations (Edwards et al., 2015). So, an old gene divergence can be either due to a deep population divergence of a small population, or to a recent divergence of a much larger population (across which individual genes can have deep coalescence times). Again it is important to note here that any difference in gene trees could be a result of the stochastic coalescent
process (i.e. the same demographic history can produce many different gene trees). This realisation led to the development of formal statistical coalescent-based models for comparative phylogeographic inference, which are capable of joint estimation of both divergence times and other parameters, such as migration rates and population sizes (Hickerson et al., 2010). Given that gene trees are largely stochastic, the ability to infer these parameters from single locus data is limited, though it can be maximised by using data from a large number of species (Stone et al., 2012; Page and Hughes, 2014; Burbrink et al., 2016; Myers et al., 2017). There has been a growing realisation that power to infer population history comes from comparing patterns across loci (Felsenstein, 2006), and so comparative phylogeographic studies are moving away from the use of single locus data and towards multi-locus analyses based on nuclear sequence or microsatellite markers (Beheregaray, 2008; Ikeda et al., 2016; Menezes et al., 2016; Pepper et al., 2017).

The rising popularity of phylogeographic studies over the last 30 years has facilitated the development of several approaches that aim to compare phylogeographic histories across species. One approach, which attempts to analyse multitaxa datasets under a unified statistical analysis, uses hierarchical Bayesian models that can be incorporated into an ABC (approximate Bayesian computation) framework (hABC) (Hickerson et al., 2006). The benefit of ABC based methods is computational tractability, as they do not calculate full likelihoods from the data, which allows complex demographic scenarios with large numbers of parameters to be investigated (Sunnaker et al., 2013). ABC simulates the genetic patterns that would be observed under a defined set of alternative population histories, and identifies the history whose genetic outcome most closely matches that seen in a set of real data. The approach involves very large numbers (~ millions) of simulations for each alternative history for molecular markers of the same length and mutational model as those sampled from the real populations. Each simulation draws population demographic parameters (population sizes and rates of change, divergence times, migration rates, mutation rates) from specified prior distributions. A set of summary statistics is then calculated for each simulation that captures genetic patterns within and between populations. These summary statistics are used to filter the simulation replicates, retaining only the subset (often 1%) that are closest to those calculated for the real sampled data in order to infer the history that best fits the observed values (Bertorelle et al., 2010). hABC is a comparative approach that adds additional cross-species hyperparameters to the analysis (Hickerson et al., 2006). For example, hABC has been used to assess the evidence that a group of n
species sampled from the same pair of habitats showed population divergence at
the same time in the past, or over 2 to \( n \) different divergence times, and estimates
the age of each (multi) species divergence pulse (Leaché et al., 2007; Barber and
Klicka, 2010; Daza et al., 2010; Lawson, 2010; Myers et al., 2017). A strength of this
approach is that the power to accept or reject specific population history models
comes from the number of species in the analysis – so it makes the most of single
locus data (Chan et al., 2014). Stone et al (2012) used a hABC approach
(implemented in the software msBayes (Hickerson et al., 2007)) to assess the support
for two different community assembly models in 31 species (gallwasps and
parasitoids) of the western Palaearctic oak gall wasp community. The study used a
single mtDNA locus to analyse sets of neighbouring population pairs. The power in
this single locus study was achieved by analysing multiple species together e.g. one
pairwise analysis comprised mtDNA from 12 gallwasp species sampled from two
populations. Although able to infer the number and age of divergence events
experienced across the sampled species, the study was unable to say which species
diverged at which time, because of the hyperprior structure of the msBayes analysis
(Hickerson et al., 2007). To keep the analysis tractable, species are allocated to co-
diverging groups in each simulation without recording species identity. This is not
an inherent constraint of hABC, but one imposed in msBayes to avoid the massive
increase in computational time that would result if large numbers of simulations
were performed for each possible set of species combinations in a large sample.

These limitations notwithstanding, the hABC approach was extended further to
allow multi-locus analysis in multiple species pairs and is implemented in MTML-
msBayes (Huang et al., 2011, see also Oaks (2014) for a modified version). Bell et al
(2012) used this approach to assess support for the presence of one or more
vicariance events across the Black Mountain Corridor in the Australian Wet Tropics
in 5 species of rainforest frog. The study was able to reject a model of a single
vicariance event in favour of multiple events using a combined nuclear and mtDNA
dataset (1 mtDNA locus and 7-10 nuclear loci). The study also compared inference
based on the mtDNA only dataset against the multi-locus dataset and, although
both suggested multiple vicariance events, the mtDNA-only dataset lacked the
statistical power (present in the combined dataset) to estimate the number and
timing of the inferred vicariance events (Bell et al., 2012).

Scaling such analyses to genome-wide datasets, likely consisting of thousands of
SNPs, is an on going challenge that ultimately involves some level of compromise,
e.g. limited sampling of individuals and/or populations or a limit on the complexity
of models analysed. One approach that uses a hABC framework to infer shared demographic responses across multiple taxa, which is based on a summary of SNP data called the aggregate SFS (aSFS), has been developed by Xue and Hickerson (2015) (also see Xue and Hickerson (2017)). The aSFS combines multiple single-taxon SFSs to form a single multitaxon summary of genetic variation. The authors found, using simulation studies, that patterns in the aSFS allow discrimination among alternative models of population expansion, and that inference can be improved with greater numbers of taxa and loci but not necessarily with greater numbers of individuals per taxon. The authors then showed how an aSFS summary of sequence data obtained from 5 populations of threespine sticklebacks could be combined with a hABC approach to infer their co-expansion history. Although this approach scales well to large datasets in terms of number of taxa and loci, it is limited in its inference of taxon-specific demographic parameters (Xue and Hickerson, 2015), focusing instead on the level of synchronicity (or not) across the whole dataset (as with msBayes, above).

A serious limitation of these ABC based approaches is that inference is based on a set of summary statistics calculated from the data. This ultimately throws away information present in the full dataset that can lead to biased parameter estimates and poor model selection (Marjoram and Tavare, 2006; Csillery et al., 2010; Robert et al., 2011; Sunnaker et al., 2013). ABC analyses require large sample sizes (with the exception of genomic datasets, see Robinson et al. (2014)) in order to gain the power necessary to statistically infer one history over another, an important consideration when generating data for comparative phylogeographic inference (Stocks et al., 2014). Currently there is no method that uses genome-wide per individual sampling to jointly estimate single-species demographic parameters and shared histories across taxa under a unified statistical framework. An alternative approach to ABC based methods, used here, is to estimate demographic parameters per species from genome-wide datasets, and then to make post-hoc comparisons of parameter values across species (Leache et al., 2013; Demos et al., 2015; Oswald et al., 2017; Satler and Carstens, 2017). This approach utilises the information present in the full dataset and is suitable for use with small sample sizes. Single species histories can be inferred using a range of demographic inference approaches, some of which are discussed in Chapters 2 and 3. However, to be able to assess whether differences in parameters across species are significant, robust estimates of the variance around each species estimate must be calculated (e.g. by parametric bootstrapping (Lohse et al., 2016; Oswald et al., 2017)).
4.2.2 Model systems for community studies

Herbivorous insects use plant tissue as a food source. They differ from other animal herbivores in distinct ways, including small individual size, often large populations, and the many unique ways they interact with their host plants. Insect herbivores tend to be more specialised in their choices of host compared to other herbivore taxa (Stamp, 2001), some show lifelong associations with their host plants, and due to their short generation time they exhibit rapid rates of evolution (Strauss and Zangerl, 2002). Plants, herbivorous insects, and their insect natural enemies comprise over half of the species on Earth (Leppanen et al., 2013), and their interactions dominate terrestrial ecosystems. Working on these communities can help to answer a huge array of ecological and evolutionary questions, and plant-insect interactions have been major subjects of research for many years.

Several systems have become particularly well-studied model examples of plant-insect interactions. These tend to be based around the feeding habits of the herbivore, and represent natural microcosms of interacting species focused on specific niches - for example, leaf miners, stem borers, and gallers (Hayward and Stone, 2005; Singer and Stireman, 2005). These systems all have a third trophic level as part of their community structure, which adds to their complexity as well as their usefulness as model systems in multi-trophic studies. This extra layer is most commonly comprised of insect parasitoids, which at a species level can be generalist or specialist in their choice of herbivore prey. While some communities are characterised by many generalist parasitoid species (for example, oak cynipid galls (Stone et al., 2002)), others have large numbers of more specialist species that target a specific subset of available herbivore hosts, resulting in tightly-linked tri-trophic interactions (Leppanen et al., 2013). The high specificity of herbivore to host plant, and natural enemy to herbivore, in these systems makes them ‘ecologically closed’, such that they can be meaningfully considered in isolation of other species and niches. This makes them excellent model systems for community studies (Lewis et al., 2002; Stireman and Singer, 2003; Bailey et al., 2009; Bihari et al., 2011). One such system that is ideally suited to community level studies are the multi-trophic communities of wasps that are associated with the genus Ficus.

4.2.3 The trophic levels of a fig wasp community

The communities of wasps associated with fig fruits are divided broadly into pollinating and non-pollinating fig wasps. The group classified as non-pollinating fig wasps (NPFW) is comprised of different trophic levels that can be further
categorised by the feeding niches of the individual species. The first trophic level consists of herbivores, which includes the pollinating fig wasps but also other galling species that induce their own galls and feed directly upon plant material (Segar et al., 2014). The second level consists of the parasites of the herbivores. These include wasps whose larvae feed upon the larvae of herbivores (parasitoids) as well as inquilines that lay their eggs in galls induced by others, and feed on plant tissue. Though herbivores, their feeding can result in the death of the herbivore larvae within the gall (lethal inquilines) (Ronquist, 1994). The third trophic level is comprised of hyperparasitoids, the parasitoids of the parasitoids (Segar et al., 2014). The lifecycles and timing of oviposition within the different trophic levels will vary according to the development stages of the fig (Figure 1.6, Chapter 1) and the presence of the preferred prey in the case of the parasitoids/inquilines. NPFW can oviposit by entering the fig syconium (like the pollinators) or, more commonly, oviposit from the outside of the fig using long ovipositors, which vary in length according to the time of oviposition in the fig development lifecycle (Borges, 2015). Individual fig species can harbour between 3 and 30 species of NPFW (Borges, 2015) making them complex closed communities. Often many closely related species of NPFW are associated with the same fig species (Segar, 2011; Darwell, 2013). Small gallers and parasitoids are more abundant than large species, body size being another factor that has been shown to correlate with the pattern of oviposition timing seen across trophic levels (Segar et al., 2014). Large wasps that induce large galls are more likely to oviposit early in fig development, prior to the arrival of pollinators, whereas small gallers more closely match the timing of the pollinators. Parasitoids and hyper-parasitoids oviposit after the pollinators in the latest stages of fig development (Segar et al., 2014). All community members generally emerge simultaneously, indicating that larval stage duration differs across trophic levels (Borges, 2015).

The extent to which these NPFW affect the obligate mutualism between the fig trees and the pollinating fig wasps is thought to differ across trophic levels. It has been shown that large gallers capable of inducing large galls during the earliest stage of fig development have a greater effect upon the obligate mutualism (in terms of seed production and pollinator numbers) than gallers and parasitoids that attack later on in fig development (Conchou et al., 2014). These large galler species do not necessarily require pollinated flowers for oviposition (some may oviposit in the fig wall rather than the syconium) (Conchou et al., 2014), demonstrated by their ability to oviposit prior to the arrival of the pollinators, and there is evidence they are able
to make their own exit holes, side-stepping the reliance upon male pollinators (Segar et al., 2014). This suggests they can have a large detrimental effect upon the mutualism as they can maintain a lifecycle without the presence of the pollinators. Other species that rely upon the pollinators’ presence within the fig, either directly (e.g. parasitoid larvae feeding upon pollinator larvae) or indirectly (e.g. reliance on male pollinators to create exit holes out of the fig (Jansen-González et al., 2014)), are assumed to have a less marked effect upon the fig lifecycle. It has been suggested that the presence of parasitoids plays an important role in maintaining the stability of the obligate mutualism by enforcing selection pressure upon the pollinators to preferentially gall flowers at the centre of the syconium, in turn reducing their galling rates and allowing seed development in the ungalled flowers (Dunn et al., 2008). Determining the precise feeding regimes and oviposition patterns of NPFW species is extremely difficult, meaning some are inferred across whole families/subfamilies or at the genus level under the assumption they will be concordant across the species within these taxonomic groups (Conchou et al., 2014). This assumption has been shown to be untrue within one subfamily (Cruaud et al., 2011), advising caution when inferring elements of community structure heavily reliant on accurate knowledge of life history traits.

### 4.2.4 Ficus rubiginosa, Ficus obliqua and their inhabitants

*Ficus rubiginosa* and *F. obliqua* are both transitional species that can be found either as freestanding trees, hemi-epiphytes or lithophytes across their range (Dixon, 2003; Ronsted et al., 2008b). It has been suggested that both species represent radiations that show transition from the rainforest into drier habitats (Ronsted et al., 2008b). *Ficus rubiginosa* is endemic to Australia and is found along the east coast throughout Queensland and New South Wales, whereas *F. obliqua* is also widespread throughout the Pacific Islands (Figures 1.9 and 1.10, Chapter 1) (Dixon et al., 2001). Both species belong to the section *Malvanthera* (subgenus *Urostigma*) but are members of different series, *Rubiginosae* and *Obliquae*, for *F. rubiginosa* and *F. obliqua* respectively (Dixon et al., 2001; Ronsted et al., 2008b). The two species look very similar, which has caused some confusion over classifications in the past (Dixon et al., 2001; McPherson, 2005), with the leaves and figs of *F. obliqua* being described as smaller analogues of those of *F. rubiginosa* (Figure 1.8, Chapter 1) (Darwell, 2013). Both fig species also harbour very similar communities of fig wasps (Darwell, 2013) but with the inhabitants of *F. obliqua*, in general, being correspondingly smaller than those of *F. rubiginosa*. Both *F. rubiginosa* and *F. obliqua* are monoecious species whose Australian east coast distributions cover several proposed biogeographic barriers.
Population genomics of a fig wasp community

(Figure 1.3, Chapter 1) (details below), making them excellent candidates for studying the historical associations within enclosed wasp communities. *Ficus* species are popular choices for planting in parks and gardens and, as *F. rubiginosa* and *F. obliqua* are both able to grow as free standing trees as well as epiphytes and lithophytes (Dixon et al., 2001), they have become especially popular for such ornamental uses.

The pollinator of *F. rubiginosa*, *Pleistodontes imperialis* (subfamily *Agaoninae*, family *Agaonidae*), has recently been described as a complex of five species, four of which are morphologically cryptic (currently designated sp 1-5) (Darwell et al., 2014). The five species occur sympatrically throughout Queensland but only a single species (sp 1) is found across New South Wales (Figure 1.11, Chapter 1). Sp 1 is therefore the only species that is found in regions that cover the target populations of this study (i.e. North, South and Sydney) and for this reason was the pollinator sampled.

Darwell (2013) reported 21 species of NPFW associated with *F. rubiginosa* across its range, six of which are included in the current study. *Sycoscapter* long (subfamily *Sycoryctinae*, family *Pteromalidae*) is a small parasitoid of the pollinator that is distributed throughout the host fig’s range (Segar, 2011; Darwell, 2013). Six *Sycoscapter* long individuals were sampled, two from each of the target populations (i.e. North, South and Sydney). *Sycoscapter* short (subfamily *Sycoryctinae*, family *Pteromalidae*) is also a small parasitoid of the pollinator but is present as two morphologically cryptic species. The distribution of these two morphs is largely allopatric with the exception of an area of overlap around Brisbane (Figure 2.2, Chapter 2; Segar, 2011; Darwell, 2013). Two individuals of each morph were included in this study. Each *Sycoscapter* species lays its eggs at fig development stage C, i.e. after the pollinators enter the syconium (Segar et al., 2014).

*Philotrypesis* (subfamily *Sycoryctinae*, family *Pteromalidae*) are small NPFW that have been designated inquilines of the pollinator in *F. rubiginosa* (and *F. obliqua*) communities (Segar, 2011; Darwell, 2013), however other members of the genus have been classified as parasitoids (Jiang et al., 2006; Zhou et al., 2012). *Ficus rubiginosa*-associated *Philotrypesis* are present in two distinguishable yellow and black forms (Segar, 2011), each of which has recently been identified to contain two putative species, suggesting a total of four *Philotrypesis* species specific to the same fig host, *F. rubiginosa* (Darwell, 2013). Four *Philotrypesis* yellow, two each of species 1 and 2, were sampled for this study, two from the South and two from the North
respectively. *Philotrypesis* wasps lay their eggs at the same time as *Sycoscapter* wasps (fig development stage C) (Segar et al., 2014).

*Watshamiella* (subfamily *Sycoryctinae*, family *Pteromalidae*) are small NPFW that are thought to be hyper-parasitoids of *Sycoscapter* wasps (Compton et al., 2009; Segar et al., 2014). They have been shown to exploit the holes drilled through the fig wall by their hosts (Compton et al., 2009), an observation supported by a study that showed *Watshamiella* wasps lay their eggs at a later fig developmental stage (between stages C and D) compared to *Sycoscapter* individuals (Segar et al., 2014). *Ficus rubiginosa*-associated *Watshamiella* wasps are found in long and short ovipositor forms, both of which have distributions covering the entire host figs range (Darwell, 2013). Six individuals of the short ovipositor species were sampled in the present study, two from each population (i.e. North, South and Sydney).

*Herodotia* (subfamily *Epichrysomallinae*, family *Pteromalidae*) are large galling NPFW wasps that oviposit early in fig development (around stage A), which is prior to the pollinators (Segar et al., 2014). *Herodotia* wasps are found across the distribution of *F. rubiginosa* and were thought to consist of a single species (Segar, 2011) until Darwell (2013) suggested that there are actually three *Herodotia* species (sp. 1, 2 and 3) associated with *F. rubiginosa*. The results of the genome-wide SNP analysis in Chapter 2 did not support a distinction between the sp. 2 and sp. 3 samples of this study (designated as such based on the cytb phylogeny) however. Therefore four *Herodotia* individuals, chosen on the basis of sampling location, were included in this study, two from the North population and two from the South population.

*Eukobelea* (subfamily *Sycophaginae*, family *Agaonidae*) consists of small galling NPFW wasps that oviposit around the same time as the pollinators (fig development stage B) (Segar et al., 2014). *Ficus rubiginosa* associated *Eukobelea* have been found to exist as two cryptic species, one occupying a largely allopatric Queensland distribution whilst the other is predominantly found in NSW (but stretches as far as Brisbane, QLD) (Figure 2.3, Chapter 2, (Darwell, 2013)). Two individuals of each species were included in this study. As the majority of species in this Chapter are associated with *F. rubiginosa*, the host tree will only be specified for those species where this is not the case.

*Ficus obliqua* is also pollinated by wasps of the genus *Pleistodontes* and its figs harbour *Sycoscapter* parasitoids that attack these pollinators. However, due to sampling errors (see Chapter 2) these species are not included in this study. Darwell
(2013) reported 16 species of NPFW associated with *F. obliqua* across its Australian range, one of which is included in the current study. NPFW of the genus *Philotrypesis* were sampled from the North and South populations and included in the study to enable a within-guild, across host tree comparison with the *Philotrypesis* individuals from *F. rubiginosa*. As with *F. rubiginosa* associated *Philotrypesis*, those associated with *F. obliqua* are thought to be inquilines of the pollinator and are found in yellow and black forms (Darwell, 2013). Although not specifically tested, it is assumed the time of attack of *F. obliqua* associated *Philotrypesis* will closely mirror that of *F. rubiginosa* associated *Philotrypesis* (i.e. after the pollinator enters the syconium) as this timing has also been observed in a distantly related *Ficus* species (Compton, 1993; Segar, 2011).

### 4.2.5 East coast biogeography

The climatic changes experienced throughout the Cenozoic in Australia that have contributed to the gradual movement towards aridity (Martin, 2006) are likely to have affected different species in different ways. For example, the range of rainforest taxa was likely restricted during cooler, dry periods but able to expand during warmer, wetter periods whereas taxa adapted to dry, open forest habitats may have exhibited contrasting responses with range expansions during the cooler, dry periods. The mesic strip along the east coast of Australia that is maintained by the presence of the Great Dividing Range (GDR) has undergone substantial habitat changes throughout the Quaternary glacial cycles (Kershaw, 2005; Chapple et al., 2011b). Such habitat shifts and their possible effects upon east coast flora and fauna are of extreme interest in the quest to understand what processes have shaped current biodiversity. Of particular interest along the east coast of Australia are the numerous biogeographic barriers that have been implicated in the phylogeographic structuring of a range of taxa (Moritz et al., 2000; Pope et al., 2001; Schäuble and Moritz, 2001; Garrick et al., 2004; Nicholls and Austin, 2005; Brown et al., 2006; Hodges et al., 2007b; Krosch et al., 2009; Moritz et al., 2009; Rossetto et al., 2009; Edwards and Melville, 2010; Beavis et al., 2011; Garrick et al., 2012; MacQueen et al., 2012; Bryant and Fuller, 2014; Mellick et al., 2014; Bryant and Krosch, 2016). Two of the most significant are the Burdekin and St. Lawrence Gaps (discussed in Chapters 1 and 3), large dry corridors located between Brisbane and Townsville (Figure 1.3, Chapter 1). Further South around the Queensland/New South Wales border, three mountain ranges (Main, McPherson and Border) extend from the GDR out to the coast, collectively referred to as the Brisbane Valley Barrier (BVB) (Bryant and Krosch, 2016). The Brisbane valley sustains a drier climate compared to the uplands.
that surround it and is therefore thought to be an important potential barrier to dry open forest taxa (Weber et al., 2014). Between the BVB and Sydney is the Hunter Valley Barrier (HV) (Figure 1.3, Chapter 1). The HV is an area of lowland that follows the Hunter River out to the coast at Newcastle (Bryant and Krosch, 2016). Like the Brisbane valley it supports a drier climate compared to the uplands that enclose it to the North and South (Barrington Tops and Blue Mountains respectively) (Chapple et al., 2011b; Smissen et al., 2013). All four of these barriers have been implicated in playing a role in shaping species distributions and driving divergence in a range of taxa, equally however, they have all also been shown to have no effect on the ranges of other taxa (Firestone et al., 1999; James and Moritz, 2000; Schäuble and Moritz, 2001; Nicholls and Austin, 2005; Brown et al., 2006; Lucky, 2011; MacQueen et al., 2012; Burke et al., 2013; Smissen et al., 2013; Bryant and Fuller, 2014) (see Chapter 1). These differing responses are inferred across both closed and open forest taxa, highlighting the difficulties in making generalisations about the joint affects of such barriers on different taxa, even those who share similar habitats (Chapple et al., 2011a; Chapple et al., 2011b; Bryant and Fuller, 2014; Weber et al., 2014; Bryant and Krosch, 2016).

4.2.6 Objectives

The overall aim of this chapter is to assess spatial patterns in the structure of the fig wasp community associated with *F. rubiginosa* along the east coast of Australia by comparing the individual population histories inferred for each of the seven species sampled. First I ask general questions about the impact of putative geographic barriers on population structure across fig-associated insects: Is there evidence of genetic divergence in populations sampled either side of each divide? Is there evidence of migration across the divide, and if so, which direction predominates? Is the direction of inferred migration shared across species? Are there differences in population size in the North and South populations, and are these consistent across species? If population sizes and migration rates are indicative of fundamental properties of the landscape (habitat patch sizes, or prevailing wind direction, for example), we might expect concordant patterns in these parameters across species. It is important to consider the fact that the patterns we might see are likely to reflect the location of the ancestral populations of the sample species. For example, was the North or South population recently ancestral to the other, or do the sampled populations reflect remnants of a wider and previously continuous range that has become subdivided due to the impact of long-term climate change on habitats? The sampling here does not allow inference of the wider history of each species (this
would require wider sampling of a larger number of sites). However, community structuring through directional processes or vicariance make different predictions for patterns in population history across species, which I address in two hypotheses (outlined below).

Second, I address questions linked to community assembly. One assembly model (directional assembly) is based on the logic that hyperparasitoids cannot exist in a location without their parasitoid hosts, and parasitoids cannot exist without their herbivore hosts. Thus in a directional model, higher trophic levels cannot arrive before lower trophic levels. If figs spread along the east coast of Australia in a particular direction (and most probably north to south as it is thought figs arrived in Australia from the Pacific Islands), then we might expect divergence times across biogeographical gaps to reflect the arrival order of species in the environment, i.e. figs ≥ herbivorous fig wasps ≥ parasitoids ≥ hyperparasitoids. If this model is true, I expect to find overall concordance in the divergence histories inferred across the *F. rubiginosa* community but different population split times. I expect the parasitoids (*Sycoscapter*), inquiline (*Philotrypesis*) and small galler (*Eukobelea*) to have younger divergence times compared to the pollinator (*Pleistodontes*), as all four rely upon the pollinator for the completion of their lifecycles and I expect the hyper-parasitoid (*Watshamiella*) to have a younger divergence time compared to its parasitoid host (*Sycoscapter*). The evidence that the large galler (*Herodotia*) is able to complete its lifecycle without the pollinator present suggests its history may be independent of the other community members.

A second model envisages community assembly prior to the appearance of the geographic barriers, such that population divergence reflects vicariance. In this model, all species are assumed to have had ancestral populations distributed right along the east coast of Australia. Climatic changes resulted in the appearance of habitat gaps, and initiated genetic divergence between populations either side of growing barriers to dispersal. Under this scenario, we expect oldest divergence, and least gene flow, in those species that are least able to disperse. Not a lot is known about the dispersal abilities of NPFW but there is evidence that pollinators of monoecious fig trees can disperse long distances on the wind. Presumably this may be the case for the NPFW also, especially as some are known to have longer adult life spans compared to the pollinators.

Additionally, I will apply the questions above to a comparison between the population histories inferred for two species from the inquiline genus, *Philotrypesis*,...
associated with *F. rubiginosa* and *F. obliqua* respectively. Based on the outwardly comparable ecologies and wasp communities of these two fig trees, I expect to find concordance between the inferred histories of the *Philotrypesis* species sampled, under the assumption that a habitat shift able to shape the history of one species may well have had a similar affect upon the other.

All population histories were inferred using genome-wide datasets, created from two haploid male wasps per sampled population, and a novel composite likelihood demographic inference method. For each species sampled, I 1) infer a demographic history for a pair of populations using a series of divergence models with different modes of gene flow, 2) assess the significance of the best fit model using the model selection process described in the methods, and, 3) obtain confidence intervals for the inferred parameters using a parametric bootstrap approach. I then compare and contrast these histories across and within species and trophic levels to address the questions above.
4.3 Methods

4.3.1 Pre-processing

VCF and BED files were generated through the pipeline outlined in Chapter 2. The VCF files contain all the filtered variant sites called across all the individuals of each species. The BED files detail the positions of the callable regions in each contig and the length of each region.

4.3.1.1 Block cutting

I cut the genomic data into aligned sequence blocks of a fixed length. Blocks were cut from homologous regions aligned across all individuals per species (Philotrypesis and Herodotia blocks were cut from aligned regions across the North and South individuals only (FRN_310_PHIL, FRN_407_PHIL, FRS_40_PHIL, FRS_247_PHIL and FRN_349_HERO, FRN_372_HERO, FRS_244_HERO, FRS_517_HERO). The average pairwise π across individuals was used to determine the length of these sequence blocks. I chose to aim for 1.5 variant sites per block. Dividing this by the average pairwise π (across all individuals) per species gave me the block lengths for each. Block lengths were fixed by the average number of variable sites to make the datasets comparable across species (block lengths are detailed in Table 4.3).

Blocks were first cut for each species (and within each species, each population pair) comparison using the block cutting procedure outlined in Chapter 3. This allowed some useful data summaries to be calculated for each species/population pair. Blocks, of the same length per species, were then cut using pairwise data summaries to enable them to be used with a different demographic inference setup. I prepared these sequence blocks using an updated version of the custom python (Python 2) script used in Chapter 3. The script summarises the blocks into the four mutational types (Ahets, Bhets, Fixed, Shared) possible when considering four haploid (or two diploid) individuals (as detailed in Chapter 3). This enables the blocks that fail the 4-gamete test (i.e. those that include both fixed differences and shared heterozygous sites) and so likely violate the assumption of no recombination within blocks to be excluded from further processing. Next, the filtered blockwise data is summarised in terms of six distributions of pairwise differences: 1) pairwise differences in population A (inclusive of any shared with population B), 2) pairwise differences in population B (inclusive of any shared with population A), 3) pairwise differences between individual 1 of population A and B, 4) pairwise differences between individual 1 of population A and individual 2 of population B, 5) pairwise differences between individual 2 of population A and individual 1 of population B,
6) pairwise differences between individual 2 of population A and B (Adiffs, Bdiffs, A1B1diffs, A1B2diffs, A2B1diffs, A2B2diffs). Blockwise counts of mutations were exported as a text file. The script relies upon `pyvcf` (version 0.6.7) and `pysam` (version 0.8.4) to process the VCF files.

### 4.3.1.2 Data summaries, k-max values and configuration counts

The text files containing the block vectors were read into `Mathematica` (Wolfram Research, Inc., Mathematica, Version 10.4, Champaign, IL (2016)) notebooks per species. Several summaries were calculated from the blockwise counts of the four site types, as in Chapter 3. The pairwise distributions were read into the corresponding per species `Mathematica` notebooks and a modified version of the function `FindkMax` was used to choose k-max values for each pairwise summary. For each pairwise summary the number of blocks containing each value of sites up to the k-max value was calculated using the function `configCounts`. The average of these counts was calculated across the four ABhet categories (i.e. the count averages of A1B1, A1B2, A2B1, A2B2) and an array containing counts of the three mutation types across the dataset was created (Ahets, Bhets, ABhets) (i.e. each dataset was summarised as three pairwise distributions of mutation counts). The k-max values used for each species are detailed in Table 4.1.
<table>
<thead>
<tr>
<th>Species</th>
<th>$k$-max values</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pleistodontes nigriventris</em> (ex <em>F. watkinsiana</em>)</td>
<td>{4, 4, 5}</td>
</tr>
<tr>
<td><em>Pleistodontes imperialis</em> (ex <em>F. rubiginosa</em>)</td>
<td>{4, 5, 6}</td>
</tr>
<tr>
<td><em>Sycoscapter long</em> (ex <em>F. rubiginosa</em>)</td>
<td>{6, 5, 7}</td>
</tr>
<tr>
<td><em>Sycoscapter short</em> (ex <em>F. rubiginosa</em>)</td>
<td>{3, 4, 7}</td>
</tr>
<tr>
<td><em>Philotrypesis</em> (ex <em>F. rubiginosa</em>)</td>
<td>{3, 2, 7}</td>
</tr>
<tr>
<td><em>Watshamiella</em> (ex <em>F. rubiginosa</em>)</td>
<td>{3, 3, 7}</td>
</tr>
<tr>
<td><em>Herodotia</em> (ex <em>F. rubiginosa</em>)</td>
<td>{5, 5, 6}</td>
</tr>
<tr>
<td><em>Eukobelea</em> (ex <em>F. rubiginosa</em>)</td>
<td>{2, 2, 6}</td>
</tr>
<tr>
<td><em>Philotrypesis</em> (ex <em>F. obliqua</em>)</td>
<td>{2, 3, 6}</td>
</tr>
</tbody>
</table>

Table 4.1: $k$-max values used during the model fitting process for each species.

4.3.2 Model fitting
4.3.2.1 Composite likelihood analyses of divergence models
A modified version of the analytic likelihood calculation used in Chapter 3 was used to fit a series of divergence scenarios for each population pair across all species. This modified approach was implemented as it represents a significant improvement in computation time (i.e. minutes compared to days for the most complex models under the bSFS scheme implemented in Chapter 3). This composite likelihood method infers population histories using three pairwise distributions generated from the data. Treating each distribution as independent greatly simplifies the likelihood calculation. The generating function for the pairwise distribution of coalescence times is used to calculate the distribution of pairwise differences for a particular model and pairwise sample. This can be interpreted as the model likelihood, given the data. The parameter estimates that maximise the composite likelihood across all distributions of pairwise differences are computed. As each block contributes to all 6 pairwise counts, the log likelihood of each model was rescaled by a factor of 6 to account for this pseudo replication in the sampling of blocks. Costa and Wilkinson-Herbots (2017) used the same pairwise distributions...
to infer the demographic history of a pair of closely related *Drosophila* species and estimated the parameters of the model by maximising the composite likelihood across all three distributions. The same range and complexity of divergence models, as detailed in Chapter 3, can be inferred using this composite likelihood method (IM and ADM, Tables 3.1 and 3.2, Figure 3.4, Chapter 3). Summarising the data as three pairwise distributions greatly eases the computational and time burden of the model fitting process but sacrifices some of the information contained in the linked sites within blocks (as used in Chapter 3). To explore this trade off between power and efficiency, a comparison between the full likelihood method used in Chapter 3 and the composite likelihood method used in this chapter will constitute the first section of the results.

### 4.3.2.2 Scaling parameter estimates

The pairwise composite likelihood calculation makes all of the same assumptions, i.e. the infinite sites mutation model, no intralocus and free interlocus recombination, the same mutation rate (\(\mu\)) across all blocks, as the full bSFS scheme does. Again, the *Drosophila melanogaster* mutation rate estimated by Keightley et al (2014) of \(2.8 \times 10^{-9}\) was used to scale the parameter estimates. As with *P. nigriventris*, the generation times of the species included in this study are not known. Based on the different lifecycles of the trophic levels of the species included, I will assume 4 generations per year for all pollinators (*Pleistodontes*) and gallers (*Herodotia* and *Eukobelea*) and 6 per year for all parasitoids (*Sycosapter* and *Wathamiella*) and inquilines (*Philotrypesis*) (J. Cook, personal communication). These assumptions and their possible impact on the results will be discussed in the Discussion.

Parameter estimates were scaled in the same manner as before (Chapter 3). The value of \(\theta\) inferred by the models is a per block estimate and so is divided by the block length for that species to obtain a per site \(\theta_{bp}\). \(\text{Na}(\text{ancestral } \text{Ne})\) is calculated as \(\text{Na} = \frac{\theta_{bp}}{4\mu}\) and the scaled \(\text{Ne}\) is calculated using \((1/B)\text{Na}\). The divergence time (in generations) in the IM models is calculated as \(T*2\text{Ne}\) and in the ADM models as \((T1+T2)*2\text{Ne}\) (the time of the admixture event is \(T1*2\text{Ne}\)). The divergence time in generations is divided by the generation time for that species to give an estimate in years.

### 4.3.2.3 Bootstrapping – correction for LD

Since the blockwise likelihood analyses ignore linkage between blocks, parametric bootstrapping was carried out to account for the effects of linkage disequilibrium in the data and obtain 95% confidence intervals (C.I) of parameter estimates. Due to
the demanding nature of the parametric bootstrap procedure (in terms of simulation and computation time), I focused on the comparison of North/South population pairs in each species. This decision is justified by the reasons presented in the results section.

*msprime* (version 0.3.1 (IM) and 0.4.0 (ADM) (Kelleher et al., 2016)) was used to generate 100 bootstrap datasets corresponding to the best-fit model under each scenario (IM and ADM). I used a python (Python 2) script (adapted from the one used in Chapter 3) to specify scaled parameters and the recombination rate for these simulations. Because the number of long contigs suitable for estimating a recombination rate varied substantially between all species, I assumed the same rate of \(2.719 \times 10^{-10}\) per bp and generation for all species (Chapter 3). This assumption and its possible affects will be addressed in the discussion. The size and number of blocks matched that of the observed data and was simulated over 20 contiguous stretches of sequence per dataset. MLE of parameters were obtained for each simulated dataset in *Mathematica* and 95% C.I. were estimated as 1.96 standard deviations (SD), either side of the MLEs.

### 4.3.2.4 Model selection

The process described in Chapter 3 was used to assess whether the better fit of the ADM model observed in the data (where applicable) was significant. The \(\Delta \ln L\) were plotted as a histogram and a critical value was obtained from the distribution (referred to as the LRT distribution in the Results) at a significance level of 0.05 using the ‘Quantile’ function in *Mathematica*. If the \(\Delta \ln L\) obtained from the data was greater than this critical value, the ADM model was deemed to fit significantly better than the IM model. This procedure was also used to establish the best-fit model for species where the addition of gene flow did not significantly (i.e. where the 95% C.I. for \(M (or f)\) overlapped zero) improve model fit and to test whether a model of divergence fitted significantly better than a model of panmixia (i.e. all four individuals are from the same population).
4.4 Results

4.4.1 Full bSFS vs pairwise composite likelihood – *P. nigriventris*

Overall the results generated using the full and composite likelihood schemes were highly concordant. Both methods inferred the same best-fit IM and ADM model for the two *P. nigriventris* populations. The predominant gene flow direction (South to North) and the asymmetry in $N_e$ (South larger than North) were consistent across both methods and models (Figure 4.1). The scaled point estimates for the divergence times were also very similar across both methods and models: the full versus the composite scheme estimates for the IM and ADM models respectively were $\sim 177$ kya (thousand years ago) and $\sim 220$ kya, and $\sim 196$ kya and 220 kya (Table 4.2). On the whole, the scaled estimates of the population $N_e s$ were comparable across methods. The largest difference between the two methods was in the estimate of the $N_e$ of the North population: under the IM model the MLEs for this parameter were $\sim 58,000$ and $\sim 41,000$ for the full and the composite scheme respectively (Table 4.2). The scaled migration rate inferred under the composite scheme was about 3 times that inferred under the full scheme whereas the admixture fraction was around 1.5 times larger under the composite compared to the full scheme (Table 4.2). Under the full likelihood method the ADM model was a significantly better fit to the data compared to the IM model. However, when the same model selection process was performed for the composite model selection, there was no significant support for one mode of gene flow over the other. The $\Delta \ln L$ between the composite ADM and IM models was -6.934 and the critical value obtained from the simulated distribution was 8.2 (p=0.05) (Table 4.2).
Figure 4.1: Diagrammatic representation of the best supported IM (left) and ADM (right) models under the full and composite likelihood schemes. Each is scaled according to the estimated divergence time in years (95% C.I. are represented by the thin black lines) and the relative difference in \(Ne\) between the two populations.
<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>$\theta$ (Na)</th>
<th>B (Ne)</th>
<th>Div* time (T) - generations</th>
<th>Div* time (T) - years</th>
<th>Adm^ time (T1) - generations</th>
<th>Adm^ time (T1) - years</th>
<th>Scaled migration rate (M)</th>
<th>Adm^ fraction (f)</th>
<th>$\Delta$lnL**</th>
<th>Critical value**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>IM2NeB A(a)</td>
<td>69,029 (65,414-72,643)</td>
<td>57,871 (54,285-61,457)</td>
<td>709,135 (689,995-728,275)</td>
<td>177,284 (172,499-182,069)</td>
<td>N/A</td>
<td>N/A</td>
<td>2.573*10^-7</td>
<td>1.639<em>10^-7 (3.506</em>10^-7)</td>
<td>4174.6</td>
<td>155.4</td>
</tr>
<tr>
<td></td>
<td>ADM2Ne BA(a)</td>
<td>69,321 (66,059-72,583)</td>
<td>58,690 (55,212-62,167)</td>
<td>782,103 (771,061-793,146)</td>
<td>195,526 (192,765-198,286)</td>
<td>229,554 (212,752-246,356)</td>
<td>57,388 (53,188-61,589)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.239</td>
<td>(0.206-0.272)</td>
</tr>
<tr>
<td>Composite</td>
<td>IM2NeB A(a)</td>
<td>71,957 (68,780-75,133)</td>
<td>41,169 (36,906-45,432)</td>
<td>879,251 (851,320-907,181)</td>
<td>219,813 (212,829-226,795)</td>
<td>N/A</td>
<td>N/A</td>
<td>7.248<em>10^-7 (6.134</em>10^-7)</td>
<td>8.363*10^-7</td>
<td>-6.934</td>
<td>8.197</td>
</tr>
<tr>
<td></td>
<td>ADM2Ne BA(a)</td>
<td>72,239 (69,210-75,267)</td>
<td>54,413 (50,269-58,557)</td>
<td>892,158 (856,430-902,071)</td>
<td>223,039 (217,334-228,744)</td>
<td>223,293 (198,742-247,844)</td>
<td>55,823 (49,685-61,961)</td>
<td>N/A</td>
<td>N/A</td>
<td>0.364</td>
<td>(0.321-0.406)</td>
</tr>
</tbody>
</table>

Table 4.2: Details of the scaled parameter MLEs for the best supported IM and ADM models under the full (top two rows) and composite (bottom two rows) schemes with their corresponding 95% C.I. in brackets. *Divergence. ^Admixture. **the difference in lnL between the best supported IM and ADM models and the critical value (p=0.05) obtained from the LRT distributions.
4.4.2 Population pair data summaries

Unless otherwise noted, the main focus throughout the results will be on the North/South population pairs. The block lengths are inversely proportional to the per species pairwise π values. Therefore the species with the lowest overall diversity has the longest blocks (P. imperialis has a per species π of 0.0013 and a block length of 1133) whereas the species with the highest overall diversity has the shortest blocks (Eukobelea has a per species π of 0.0155 and a block length of 97) (Table 4.3).

As a consequence, the number of blocks per species is not only a result of the total amount of data but also of the block length (e.g. post-filtering, the lowest number of blocks was 7,532 in P. imperialis and the largest was 982,643 in Eukobelea). Table 4.3 shows the percentages of blocks lost through the 4-gamete filtering. Diversity within and divergence between-populations was assessed using pairwise π.

Unsurprisingly, between-population π was highly concordant across each of the 4 pairwise comparisons per species but highly variable across species (Figure 4.2). Three species (P. nigroviventris (ex F. watkinsiana), Sycoscaptar long and Herodotia) showed very little difference between North and South π. In five species (P. imperialis, Sycoscaptar short, Philotrepsis (ex F. obliqua), Watshamiella and Eukobelea) the South population was more diverse than the North, while in one species (Philotrepsis) the North population showed the higher diversity. These summaries, along with Fst and intra-population heterozygosities (Table 4.3), gave crucial information about each species, guiding the model fitting process.
<table>
<thead>
<tr>
<th>Species</th>
<th>Per species $\pi^*$</th>
<th>Block length</th>
<th>Pop pair</th>
<th>Percentage of blocks lost to filtering</th>
<th>Dataset</th>
<th>No. blocks</th>
<th>{A,B,F,S} proportions^</th>
<th>Pop A/Pop B heterozygosities</th>
<th>Fst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleistodontes nigriventris (pollinator)</td>
<td>0.00388</td>
<td>387</td>
<td>N/S</td>
<td>0.76%</td>
<td>Unfiltered</td>
<td>775,997</td>
<td>{0.164,0.153,0.673,0.010}</td>
<td>0.000953/0.000889</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>770,105</td>
<td>{0.161,0.149,0.685,0.005}</td>
<td>0.000884/0.000822</td>
<td>0.74</td>
</tr>
<tr>
<td>Pleistodontes imperialis (pollinator)</td>
<td>0.00132</td>
<td>1133</td>
<td>N/S</td>
<td>1.52%</td>
<td>Unfiltered</td>
<td>7,648</td>
<td>{0.120,0.416,0.448,0.015}</td>
<td>0.000377/0.0012</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>7,532</td>
<td>{0.112,0.419,0.463,0.005}</td>
<td>0.000308/0.0011</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N/SYD</td>
<td>1.48%</td>
<td>Unfiltered</td>
<td>7,648</td>
<td>{0.124,0.399,0.465,0.012}</td>
<td>0.000378/0.00114</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>7,535</td>
<td>{0.114,0.403,0.480,0.003}</td>
<td>0.000307/0.00107</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S/SYD</td>
<td>3.65%</td>
<td>Unfiltered</td>
<td>7,648</td>
<td>{0.406,0.379,0.075,0.139}</td>
<td>0.00120/0.00114</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>7,369</td>
<td>{0.419,0.394,0.058,0.129}</td>
<td>0.00108/0.00103</td>
<td>0</td>
</tr>
<tr>
<td>Sycoscapter long (parasitoid)</td>
<td>0.00159</td>
<td>938</td>
<td>N/S</td>
<td>6.39%</td>
<td>Unfiltered</td>
<td>11,182</td>
<td>{0.342,0.355,0.160,0.143}</td>
<td>0.00153/0.00157</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>10,468</td>
<td>{0.361,0.378,0.150,0.111}</td>
<td>0.00123/0.00127</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N/SYD</td>
<td>5.85%</td>
<td>Unfiltered</td>
<td>11,182</td>
<td>{0.345,0.363,0.154,0.138}</td>
<td>0.00153/0.00159</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>10,528</td>
<td>{0.360,0.385,0.144,0.111}</td>
<td>0.00126/0.00132</td>
<td>0.11</td>
</tr>
<tr>
<td>Species</td>
<td>Per species π*</td>
<td>Block length</td>
<td>Pop pair</td>
<td>Percentage of blocks lost to filtering</td>
<td>Dataset</td>
<td>No. blocks</td>
<td>{A,B,F,S} proportions^</td>
<td>Pop A/Pop B heterozygosities</td>
<td>Fst</td>
</tr>
<tr>
<td>------------------------------</td>
<td>----------------</td>
<td>--------------</td>
<td>----------</td>
<td>----------------------------------------</td>
<td>------------</td>
<td>------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>------</td>
</tr>
<tr>
<td><em>Sycoscapter short</em> (parasitoid)</td>
<td>0.00933</td>
<td>161</td>
<td>N/S</td>
<td>0.24%</td>
<td>Unfiltered</td>
<td>760,963</td>
<td>{0.126,0.202,0.667,0.005}</td>
<td>0.00184/0.00291</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>759,155</td>
<td>{0.126,0.201,0.669,0.004}</td>
<td>0.00181/0.00287</td>
<td>0.73</td>
</tr>
<tr>
<td><em>Philotryptesis</em> (inquiline)</td>
<td>0.01268</td>
<td>118</td>
<td>N/S</td>
<td>0.06%</td>
<td>Unfiltered</td>
<td>601,744</td>
<td>{0.083,0.044,0.872,0.0006}</td>
<td>0.00163/0.00088</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>601,374</td>
<td>{0.083,0.044,0.873,0.0004}</td>
<td>0.00162/0.00087</td>
<td>0.90</td>
</tr>
<tr>
<td><em>Watshamiella</em> (hyper-parasitoid)</td>
<td>0.00767</td>
<td>195</td>
<td>N/S</td>
<td>0.14%</td>
<td>Unfiltered</td>
<td>536,301</td>
<td>{0.081,0.136,0.777,0.007}</td>
<td>0.00116/0.00189</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>535,549</td>
<td>{0.080,0.136,0.778,0.006}</td>
<td>0.00114/0.00187</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N/SYD</td>
<td>0.14%</td>
<td>Unfiltered</td>
<td>536,301</td>
<td>{0.081,0.138,0.775,0.006}</td>
<td>0.00117/0.00193</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>535,544</td>
<td>{0.080,0.138,0.776,0.006}</td>
<td>0.00114/0.00187</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S/SYD</td>
<td>0.17%</td>
<td>Unfiltered</td>
<td>536,301</td>
<td>{0.411,0.419,0.061,0.109}</td>
<td>0.00193/0.00195</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered</td>
<td>535,415</td>
<td>{0.413,0.421,0.058,0.107}</td>
<td>0.00191/0.00194</td>
<td>0.006</td>
</tr>
<tr>
<td>Species</td>
<td>Per species $\pi^*$</td>
<td>Block length</td>
<td>Pop pair</td>
<td>Percentage of blocks lost to filtering</td>
<td>Dataset</td>
<td>No. blocks</td>
<td>{A,B,F,S} proportions$^\wedge$</td>
<td>Pop A/Pop B heterozygosities</td>
<td>Fst</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>----------</td>
<td>---------------------------------------</td>
<td>----------</td>
<td>------------</td>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td><strong>Herodotia</strong> (galler)</td>
<td>0.00236</td>
<td>633</td>
<td>N/S</td>
<td>3.60%</td>
<td>Unfiltered 14,459</td>
<td>{0.386,0.383,0.128,0.103}</td>
<td>0.00248/0.00246</td>
<td>0.094</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered 13,938</td>
<td>{0.397,0.392,0.119,0.091}</td>
<td>0.00231/0.00228</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td><strong>Eukobelea</strong> (galler)</td>
<td>0.01546</td>
<td>97</td>
<td>N/S</td>
<td>0.09%</td>
<td>Unfiltered 983,490</td>
<td>{0.068,0.108,0.823,0.0006}</td>
<td>0.00167/0.00266</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered 982,643</td>
<td>{0.068,0.108,0.824,0.0002}</td>
<td>0.00166/0.00265</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td><strong>Philotrypesis</strong> (ex F. obliqua) (inquiline)</td>
<td>0.01399</td>
<td>107</td>
<td>N/S</td>
<td>0.08%</td>
<td>Unfiltered 844,539</td>
<td>{0.046,0.106,0.847,0.0006}</td>
<td>0.00112/0.00256</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Filtered 843,849</td>
<td>{0.046,0.106,0.848,0.0002}</td>
<td>0.00110/0.00255</td>
<td>0.88</td>
<td></td>
</tr>
</tbody>
</table>

*Table 4.3:* Comparison of the blockwise datasets pre and post filtering. $^*$Per species $\pi$ calculated as an average across each pairwise $\pi$ value. $^\wedge$Proportions of Ahets, Bhets, Fixed and Shared sites.
Figure 4.2: Pairwise diversity within (right) and divergence between-populations (left).

4.4.3 Composite likelihood analyses

Data for each species were fitted to a series of nested IM and ADM models (see Chapter 3, Tables 3.1 and 3.2). The best-fit model for each was determined using both parametric bootstrapping and the model selection process described in the methods (LRT distributions). I first outline this comparison across all sampled species and highlight particular cases that were less straight-forward (*P. imperialis, Herodotia, Sycoscapter long and Sycoscapter short*). Finally, I assess the demographic histories in more detail for one trophic level (*Philotrypes*) and conduct a community-wide comparison of the *F. rubiginosa* species.

The ADM model, with significant population divergence and South to North gene flow, fitted significantly better than the IM model in four species (*Philotrypes*, *Watshamiella*, *Eukobelea* and *Philotrypes* (ex *F. obliqua*) (Table 4.4). The four other species showed greatest support for different models, lacking evidence for gene flow and/or population divergence between North and South populations. In *P. imperialis*, the data supported a strict divergence model with different *Ne* parameters.
in the North and the South, with no improvement in model fit from the addition of either discrete or continuous gene flow. The best-fit model (STRDIV2Ne(a)) was confirmed by a parametric bootstrap analysis, showing the $T$ parameter does not overlap zero, and by comparing the $\Delta \ln L$ to a critical value obtained from the LRT distribution (Table 4.5). For *Herodotia*, the data supported strict divergence with a single $Ne$ parameter, without gene flow (discrete or continuous). The best-fit model (STRDIV1Ne) was supported by a parametric bootstrap analysis and the LRT distribution (Table 4.5). *Sycoscapter* long showed no significant divergence between any of the population pairs tested (i.e. North/South, North/Sydney, South/Sydney), suggesting that this species is a single panmictic unit across its east coast distribution. The result was confirmed for the North/South pair using a LRT distribution where no addition of parameters was significantly supported and a bootstrap analysis where the $T$ parameter overlapped zero (Table 4.5).
Table 4.4: The scaled parameter MLEs for the best supported ADM models for *Philotrypesis*, *Watshamiella*, *Eukobelea* and *Philotrypesis (ex F. obliqua)* with their corresponding 95% C.I. in brackets. *Divergence. ^Admixture. **the difference in lnL between the best supported IM and ADM models and the critical value (p=0.05) obtained from the LRT distributions.
Table 4.5: The scaled parameter MLEs for the best supported strict divergence models for *Pleistodontes imperialis* and *Herodotia* and the panmixia model for *Sycoscapter long* with their corresponding 95% C.I. in brackets. *Divergence. **the differences in lnL and the critical value (p=0.05) obtained from the LRT distributions. ^the difference between the best supported IM and STRDIV2Ne(a) models. ^^the difference between the panmixia and STRDIV1Ne models.
4.4.3.1 *Sycoscapter short*, the exception to the rule

Data for *Sycoscapter short* supported a model of population divergence with gene flow, but in contrast to other species for which gene flow was inferred, the IM (rather than the ADM) model gave the best fit. The $\Delta \ln L$ between the best-fit ADM and IM model for *Sycoscapter short* was -318.01, suggesting the IM is a better fit to the data - a result confirmed by the LRT distribution (Table 4.6). The parametric bootstrap analysis performed for the IM model however generated extremely wide 95% C.I. around the MLE for the inferred divergence time (~1.2 mya, 95% C.I. 514 kya – 1.98 mya; Table 4.6). The percentage of lineages that are expected to have been affected by a migration event prior to divergence ($e^{(T^*(M/2))}$) was calculated as 99%. Because only the remaining 1% of lineages not affected by migration contribute information about the time of divergence, the variance around the estimate of $T$ is very high. To investigate this further, the *Sycoscapter short* data were fitted to a migration-only model (Mig), which assumes an exchange of migrants at rate $M$, with asymmetric $N_{es}$, but without population divergence (i.e. the two populations are assumed to have persisted indefinitely). The $\Delta \ln L$ (10.8) between the Mig and IM models was significant (assuming a chi-squared distribution) in favour of the IM model, however given that the chi-squared distribution cannot be assumed under the composite likelihood scheme and the large variance around $T$ under the IM model, a LRT distribution was generated for the Mig model. The critical value (5.4, $p=0.05$) obtained from the LRT distribution confirms the IM model does fit significantly better than the Mig model. These results will nevertheless by interpreted keeping the wide confidence intervals around the divergence time in mind.
<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>$\theta$ ($N_a$)</th>
<th>$B$ ($N_e$)</th>
<th>Div* time (T) - generations</th>
<th>Div* time (T) - years</th>
<th>Scaled migration rate ($M$)</th>
<th>$\Delta\ln L^{**}$</th>
<th>Critical value**</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Sycoscapter</em> short</td>
<td>IM2NeAB(b)</td>
<td>158,433 (150,418-166,449)</td>
<td>112,541 (99,647-125,435)</td>
<td>7.474<em>10^6 (3.086</em>10^6-1.186*10^7)</td>
<td>1.246<em>10^6 (514,309-1.977</em>10^6)</td>
<td>5.723<em>10^-7 (5.385</em>10^-7-6.060*10^-7)</td>
<td>-318.014</td>
<td>-451.309</td>
</tr>
</tbody>
</table>

*Table 4.6: The scaled parameter MLEs for the best supported IM for *Sycoscapter* short with their corresponding 95% C.I. in brackets. Divergence. **the difference in $\ln L$ between the best supported IM and ADM models and the critical value ($p=0.05$) obtained from the LRT distribution.*
4.4.3.2 The Sydney population

The results from the models involving the Sydney population for *P. imperialis* and *Watshamiella* supported the same overall conclusion, with no detectable differentiation between the South and Sydney populations. Although not confirmed through bootstrap analyses, the North/South and North/Sydney pairs show highly concordant results and the South/Sydney pairs show no support for divergence in both species. Both results imply an absence of population structure between the South and Sydney, i.e. these regions can be thought of as a single population.

4.4.4 The *F. rubiginosa* community

4.4.4.1 Pattern of North-South divergence across trophic levels

The estimated time of divergence, where present, varied widely across the seven *F. rubiginosa* species. Diagrammatic representation of the histories inferred for each of the seven species is shown in Figure 4.3 alongside the expectations under the ‘Directional’ and ‘Vicariance’ models of community assembly introduced in section 4.2.6. Excluding *Sycoscapter* long (parasitoid), for which no population divergence was inferred, divergence times ranged from ~24 kya (95% C.I. 19 kya – 29 kya), the youngest inferred under the strict divergence model in *Herodotia* (large galler) to ~1.4 mya (95% C.I. 1.42 mya – 1.46 mya), the oldest inferred under the ADM model in *Eukobelea* (small galler) (Tables 4.4 and 4.5). Of the species where divergence is inferred, the small NPFW (*Sycoscapter* short, *Philotrypesis*, *Watshamiella* and *Eukobelea*) that rely upon the presence of the pollinator all show significantly older divergence times than the pollinator, *P. imperialis* (~34 kya (95% C.I. 28 kya – 39 kya)) and the large galling NPFW *Herodotia* (Figure 4.3). The 95% C.I. for the divergence times of *P. imperialis* and *Herodotia* overlap, suggesting they are not significantly different (Table 4.5). Estimated divergence times for the small NPFW, however, span a large time period that incorporates both the late (*Watshamiella*, ~550 kya (545 kya – 558 kya)) and the early Pleistocene (*Eukobelea*, ~1.4 mya). Given the large 95% C.I. (~514 kya – 1.98 mya) around the divergence time of *Sycoscapter* short, it cannot be classed as significantly different to the other three small NPFW species. However, the 95% C.I. around the estimates for *Watshamiella*, *Philotrypesis* and *Eukobelea* do not overlap (Figure 4.3), suggesting they are all significantly different. The small galler, *Eukobelea* shows the oldest population divergence time (~1.4 mya), followed by a parasitoid, *Sycoscapter* short (~1.2 mya), then the inquiline, *Philotrypesis* (~900 kya). The species with the most recent North/South population divergence is the hyper-parasitoid, *Watshamiella* (~550 kya). It is interesting to note that the three species that are directly reliant on the pollinator for completion of
their lifecycles all have older divergence times than the hyper-parasitoid, *Watshamiella*, which is reliant upon the parasitoid, *Sycosapter*, for the completion of its lifecycle.
<table>
<thead>
<tr>
<th>Trophic level</th>
<th>Hypotheses relating relative divergence times across trophic levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Vicariance &lt;= fig tree &gt;&gt; pollinator &lt;= fig tree &lt;= pollinator</td>
</tr>
<tr>
<td>3</td>
<td>Directional &lt;= fig tree &gt;&gt; pollinator &lt;= pollinator &lt;= pollinator</td>
</tr>
<tr>
<td>4</td>
<td>&gt;&gt; pollinator &lt;= pollinator &lt;= parasitoid &gt;&gt; pollinator</td>
</tr>
</tbody>
</table>

- **Pollinator (P. imperialis)**
- **Galler (Herodota)**
- **Parasitoid (Sycocapther longi)**
- **Parasitoid (Sycocapther shorti)**
- **Inquiline (Philoptyvesis)**
- **Galler (Eukobeleo)**
- **Hyperparasitoid (Watshamiana)**

The figure illustrates the relationships between the different trophic levels and the species involved, with arrows indicating the direction of divergence and interaction times.
Figure 4.3: Diagrammatic representations of the inferred population histories for the *F. rubiginosa*-associated wasp community. Divergence times in years are scaled relative to the timelines. The 95% C.I. of the divergence times are represented by the thin black lines. The population sizes are scaled relative to the difference between the North and South *Nes* (where applicable). The schematic for *Sycoscapter* long represents a single population. The proposed trophic level is given above each species name (note: the fig tree is assumed to be the first level). 'Vicariance' and 'Directional' relate to two hypotheses of community assembly. The Vicariance hypothesis relates to the dispersal ability of community members and whether they can maintain gene flow across barriers that separate populations. Pollinators are known to disperse long distances on the wind but not a lot is known about the dispersal ability of NPFW's. Here the assumption is that the higher trophic levels are unable to disperse as effectively as the pollinators. The Directional hypothesis states that higher trophic levels cannot arrive before lower trophic levels, i.e. each level cannot survive without its host. Given these hypotheses the arrows at the top show the expectation of the relative divergence times of each level compared to the level below it. E.g. the pollinator is expected not to have diverged before its fig tree host, i.e. it will have a divergence time younger or equal to that of its host (<=).
4.4.4.2 Presence and direction of gene flow

Post-divergence gene flow was significantly supported in four (Sycoscapter short, Philotrypesis, Watshamiella, Eukobelea) of the seven F. rubiginosa-associated species (Figure 4.3). Three of these small NPFW species (Philotrypesis, Watshamiella, Eukobelea) showed a discrete pulse of gene flow (the ADM model) in a South to North direction. Sycoscapter short (parasitoid) was the odd one out, showing continuous gene flow (the IM model) in a North to South direction. The admixture pulses in Philotrypesis, Watshamiella and Eukobelea are all estimated to have occurred at significantly different times (their 95% C.I.s do not overlap), each being separated by ~100 kya. Eukobelea shows the oldest admixture pulse (~333 kya, 95% C.I. 318 kya – 348 kya), then Philotrypesis (~208 kya, 95% C.I. 204 kya – 211 kya) and the youngest in Watshamiella (~103 kya, 95% C.I. 99 kya – 107 kya). Interestingly however, there is no significant difference between the admixture fraction (f) inferred for these three species, and each is estimated at between 0.5 and 0.6 (Table 4.4). This indicates that each species underwent an admixture event where around half of the individuals moved from the South into the North population. The addition of gene flow did not significantly improve the fit of the divergence model in the pollinator, P. imperialis and the large galler, Herodotia. No divergence event was inferred in the parasitoid, Sycoscapter long, suggesting a high rate of dispersal and a high level of gene flow across its entire range.

4.4.4.3 Ne concordance in the North and South across the community

Five histories (out of the six divergence models inferred) significantly supported an asymmetry in Ne across the North and South populations (Figure 4.3). The location of the larger (or smaller) Ne however was not concordant across all five species. The pollinator (P. imperialis), hyperparasitoid (Watshamiella) and small galler (Eukobelea) all supported a history where the North population is smaller. In contrast, the parasitoid (Sycoscapter short) and the inquiline (Philotrypesis) supported histories involving a smaller South population (Tables 4.4 and 4.6). The Ne estimates for the North population across all five species can be grouped into four different groups. The smallest Ne in the North (~23,000, 95% C.I. 16,000 – 29,000) is in the pollinator (P. imperialis), followed by the hyperparasitoid (Watshamiella), the inquiline (Philotrypesis) and small galler (Eukobelea) (no significant difference in Ne) and lastly, the largest Ne (~158,000, 95% C.I. 150,000 – 166,000) is in the parasitoid (Sycoscapter short) (Tables 4.4 – 4.6). The Ne sizes in the South across all five species can also be grouped into four significantly different sizes, the order of which conflicts with the Northern population sizes. The smallest Ne in the South (~76,000, 95% C.I. 70,000 –
83,000) is in the inquiline (*Philotrypesis*), followed by the pollinator (*P. imperialis*) and the parasitoid (*Sycoscapter short*) (no significant difference in Ne), the hyperparasitoid (*Watshamiella*), and lastly, the largest Ne (~236,000, 95% C.I. 228,000 – 244,000) is in the small galler (*Eukobelea*). Out of the three species where admixture (South into the North) was inferred (*Philotrypesis, Watshamiella and Eukobelea*), two (*Watshamiella* and *Eukobelea*) show gene flow going from the population with the larger Ne into the one with the smaller Ne whilst *Philotrypesis* shows the opposite, i.e. gene flow going from the smaller Ne into the larger Ne.

### 4.4.5 *Philotrypesis (ex F. rubiginosa) vs Philotrypesis (ex F. obliqua)*

An ADM model was inferred as the best-fit for both *Philotrypesis* species sampled (Table 4.4). The two species shared several features of their inferred histories (Figure 4.4). In both the admixture pulse was inferred as occurring from the South into the North population with admixture fractions of similar magnitudes (*Philotrypesis (ex F. rubiginosa)*, ~0.57 (95% C.I. 0.52 – 0.61), *Philotrypesis (ex F. obliqua)*, ~0.46 (95% C.I. 0.42 – 0.51)) (Table 4.4). The inferred divergence and admixture times were ~900 kya (95% C.I. – 895 kya – 906 kya) and 208 kya (95% C.I. 204 kya – 211 kya) and ~846 kya (95% C.I. 838 kya – 854 kya) and 151 kya (95% C.I. 144 kya – 158 kya) for the *F. rubiginosa* and *F. obliqua* species respectively (Figure 4.4). The histories of the two species differ in the population inferred with the smaller (or larger) Ne. *Philotrypesis (ex F. rubiginosa)* has an Ne in the North that is over twice the size of that in the South (~136,000 and ~76,000, respectively) whereas *Philotrypesis (ex F. obliqua)* has an Ne in the North that is less than half the size of that in the South (~86,000 and ~208,000, respectively) (Table 4.4).
Figure 4.4: Diagrammatic representations of the inferred population histories for the two *Philotrypesis* (ex *F. rubiginosa* (left) and ex *F. obliqua* (right)) species (both inquilines). Divergence times in years are scaled relative to the timelines. The 95% C.I. of the divergence times are represented by the thin black lines. The population sizes are scaled relative to the difference between the North and South *Nes*.
4.5 Discussion

4.5.1 Six out of seven members of the *F. rubiginosa* community diverged in the Pleistocene

Seven species, believed to be entirely specific to their host *F. rubiginosa*, were sampled to assess the level of concordance in population histories across the trophic levels of this fig wasp community. The results show interesting patterns of concordance across the different parameters estimated per species but also highlight variation in the histories inferred. A history involving divergence between the North and South populations was inferred for six out of the seven species, the timing of each divergence event being placed within the Pleistocene (~11.7 kya – 2.58 mya, (Head and Gibbard, 2015b)). The Pleistocene was a period of global climatic change that had a major impact upon the Australian flora and fauna despite the continent not being directly affected by glaciers (Hewitt, 2000; Barrows et al., 2002). The Australian mainland experienced cycles of cool, dry and warm, wet climates corresponding to glacial and interglacial periods respectively (Pepper et al., 2014). The Pleistocene divergence events inferred for the six *F. rubiginosa*-associated fig wasps in this study spanned ~24 kya to ~1.4 mya. I will initially discuss them by splitting them into two groups that easily fall out from the results: 1) the pollinator (*P. imperialis*) and large galler (*Herodotia*) whose divergence times are ~34 kya and ~24 kya, respectively, and 2) the small galler (*Eukobelea*), parasitoid (*Sycoscapter* short), inquiline (*Philotrypesis*) and hyperparasitoid (*Watshamiella*) whose divergence times are ~1.4 mya, ~1.2 mya, ~900 kya and ~550 kya, respectively.

4.5.1.1 *Pleistodontes imperialis* and *Herodotia*

The young divergence times (~19 kya – 39 kya, inclusive of the overlapping 95% C.I) of the first group cover the boundary of Marine Isotope Stages (MIS) 2 and 3 (MIS 2, ~11 kya - ~27 kya, MIS 3, ~27 kya – 60 kya) (Van Meerbeeck et al., 2009; Railsback et al., 2015). MIS 3 is characterised by abrupt warming phases that occurred in between the glacial periods of MIS 4 and MIS 2 (Siddall et al., 2008; Van Meerbeeck et al., 2009). MIS 2 was a severe glacial period that included the last glacial maximum (LGM) (~19 kya – 21 kya) (Van Meerbeeck et al., 2009). Despite the warming phases seen in MIS 3, the general pattern was one of cooling from the last major interglacial (MIS 5) (Railsback et al., 2015), suggesting that the divergence events inferred for *P. imperialis* and *Herodotia* took place during the increased cooling, and resultant drying in Australia, that led to the extreme conditions of the LGM. Both of these species were sampled from north and south of the Burdekin (BG) and St. Lawrence Gaps (StLG) (Figure 1.3, Chapter 1). However, in both cases
the two South individuals are also split from each other by the Brisbane Valley Barrier (BVB). The fact that the four North/South pairwise $\pi$ estimates (Figure 4.2) show little discrepancy across values suggests the level of variation in both South individuals is consistent with them originating from a single panmictic population and hence indicates that the BVB has not acted as a barrier to gene flow in these species. This result is consistent with several other studies that have assessed the effects of the BVB on population structure in a range of taxa (mammals e.g. MacQueen et al., 2012, birds e.g. Nicholls and Austin, 2005, reptiles e.g. Chapple et al., 2005, invertebrates e.g. Christidis and Dean, 2008, plants e.g. Burke et al., 2013). The results presented here for *P. imperialis* and *Herodotia* do provide evidence of divergence across the BG and StLG within the same timeframe. This could be indicative of a distribution shift in the host tree *F. rubiginosa* in response to the extreme changes in climate around the LGM. However this is not necessarily supported by what is known about the current distribution of *F. rubiginosa* and its ability to exist across a range of habitats. *Ficus rubiginosa* is distributed along the length of the east coast (Figure 1.9, Chapter 1) and is a habitat generalist, able to grow as a hemi-epiphyte or lithophyte as well as a free standing tree in dry open forests (Dixon, 2003; Ronsted et al., 2008b). Therefore the fact that the results support a distribution shift associated with changes in climate, predominantly the drying of the continent, is perhaps surprising in the light of the known ecology of extant host populations.

It is also worth noting some species-specific considerations when interpreting these results for *P. imperialis* and *Herodotia*. *Pleistodontes imperialis* (species 1 sampled here) is a member of a complex of five putative species of *P. imperialis* that pollinate *F. rubiginosa* across its range (Figure 1.11, Chapter 1) (Darwell et al., 2014). This likely indicates a complex (and possibly reticulate) history that the sampling here has only captured a small part of. It could be that the ancestor of these species was widespread across the range of *F. rubiginosa* and that the lineage leading to species 1 has recently diverged between the North and South (vicariance). It could also indicate a situation where the ancestor of species 1 originated in the South and the divergence signal revealed here represents a recent expansion into the North. The large discrepancy in $N_e$ between the North and South (the South is over 4 times larger) is concordant with the proposed distribution of species 1 (Figure 1.11, Chapter 1), which is found across southern Queensland and is the dominant species in New South Wales, whereas in the North it appears to be restricted to areas of northern Queensland around Cairns (Darwell et al., 2014). This asymmetry in $N_e$ is
also supported by the larger effective size of the southern population when the two North individuals were compared with the two Sydney individuals. The absence of significant gene flow between the North and South is not surprising given the distance between the two populations and the young divergence time. However, this also needs to be considered with regard to the other species present across the hosts range. *Ficus rubiginosa* is pollinated by a range of other *P. imperialis* species, removing the necessity (as far as *F. rubiginosa* fruit set is concerned) for species 1 to populate these areas. To be able to get a more complete picture of the history that has led to this complex of pollinators, extensive sampling that incorporates all species and appropriate outgroups would be required, analysed with historical inference approaches able to incorporate larger numbers of individuals and populations.

Diversity patterns in *Herodotia* suggest a lack of clear North/South population differentiation. Pairwise diversity estimates (Figure 4.2) indicate comparable divergence within and between populations and the most complex divergence models with gene flow failed to converge for this species. Although the best-fit model indicates support for a strict divergence model with distinct North/South populations over a model of panmixis, it is not clear how much faith we can have in such a model without the ability to also test more complex models. The coalescent tree under a model of young divergence and high gene flow closely resembles one of panmixia suggesting such situations can be hard to tease apart.

### 4.5.1.2 The small NPFW

#### 4.5.1.2.1 Early-Middle Pleistocene divergence

The second group of divergence times all fall within the Early-Middle Pleistocene Transition (EMPT) (~400 kya – 1.4 mya) (Head and Gibbard, 2015a). The beginning of this period marked the start of a major climate shift globally. This period covers MISs 45 – 11 and is characterised as the transition from 41 kya to 100 kya climate cycles that progressively increased in intensity and variability (Head and Gibbard, 2015a). In Australia, a gradual decrease in precipitation and a cyclical change in vegetation cover occurred during this period (Martin, 2006). Climate shifts of this nature must have favoured the most adaptable flora and fauna, with those unable to adapt quickly enough being confined to small patches of suitable habitat or lost entirely. The end of the EMPT occurs at the MIS 12/11 boundary. This marks the end of one of the most extreme glacial cycles of the Pleistocene (MIS 12) and the start of one of the most pronounced interglacials (MIS 11) (Head and Gibbard,
Population genomics of a fig wasp community

2015a). Beyond this (i.e. ~400 kya to the end of the Pleistocene ~11 kya) there is an abrupt shift in the intensity of the interglacial cycles, making them shorter and stronger than during the EMPT, causing - in Australia in particular - more severe drying than seen previously (Martin, 2006; Lang and Wolff, 2011; Head and Gibbard, 2015a).

The oldest divergence event for the small NPFW species sampled here was ~1.4 mya, inferred in the small galler, *Eukobelea*. This occurred during MIS 45 (~1.4 mya), which is an interglacial stage that marks the start of the EMPT (Lisiecki and Raymo, 2005; Head and Gibbard, 2015a). The sampling of the North and South individuals supports the hypothesis that the divergence between these two populations may have been influenced by the BG and StLGs but does not support a role of the BVB and the HV in shaping the population structure. However the fact that both species have been found either side of the BG and StLG (i.e. the predominantly southern species has been sampled in the North and the predominantly northern species spans an area including both gaps) (Darwell, 2013) suggests that the divergence event detected here is not the result of a biogeographical barrier that has maintained two distinct populations since its occurrence. It is possible the current distribution has resulted from a divergence event, driven by the climate changes at the beginning of the EMPT, which was followed by subsequent range expansion(s).

The second oldest divergence event was ~1.2 mya inferred in the parasitoid, *Sycoscapter* short. This time is placed within MIS 36 (~1.2 mya), a glacial period characterised by more severe cooling compared to previous glacials (Head and Gibbard, 2015a). However, the large 95% C.I. (~514 kya – 1.98 mya) for this divergence time estimate, make it difficult to interpret. The time period covered by the C.I. covers the majority of the EMPT and extends further back into the early Pleistocene. It is worth noting that the divergence time inferred under the less well-fitting ADM model was ~650 kya, which falls within the C.I. of the IM divergence time but is almost half the time of the point estimate (~1.2 mya). As mentioned previously, the sampling of this species did not fit the North/South sampling of the other species due to its proposed distribution (Figure 2.2, Chapter 2) (Darwell, 2013). The two northern individuals were located from either side of the BG and StLG and the two southern individuals from either side of the HV, suggesting that these biogeographic barriers have not played a role in the population structure of these sampled individuals. However, the northern and southern individuals are separated by the BVB. Therefore, the old and significant divergence estimates are consistent with the BVB contributing to population structure.
The divergence time (~900 kya) inferred for the inquiline Philotrypesis occurs within MIS 23 (Railsback et al., 2015). MIS 23 is a weak interglacial in the midst of the ‘~900 kya event’ (MIS 22 – 24, (~866 kya – 936 kya)) (Clark et al., 2006; Head and Gibbard, 2015a), the longest glacial cycles of the Pleistocene thus far. The North and South Philotrypesis individuals were sampled either side of the BG and StLG and the significant divergence shown here supports a role for these biogeographical barriers in structuring the populations of F. rubiginosa-associated Philotrypesis. However, the fact that these species have been found to occur sympatrically in locations North of the BG (Darwell, 2013) suggests a more complex picture than a simple divide across these putative barriers. It is possible a range expansion of the South population into the North regions sometime post divergence occurred and the two species have been able to co-occur in these areas since then. Further sampling from these areas of overlap could help to complete the picture.

The youngest divergence event for the small NPFW species sampled here was ~550 kya inferred in the hyperparasitoid, Watshamiella. This occurred during MIS 14 (~530 kya – 560 kya), which is a particularly weak glacial stage of similar magnitude to the interglacials (MIS 13 and 15) that surround it, all three stages of which occur between two significantly strong glacials (MIS 12 and 16) (Lang and Wolff, 2011; Head and Gibbard, 2015a). The South population individuals were sampled either side of the BVB indicating this divide is not acting as a barrier to gene flow in this population. The fact that the Sydney individuals were shown to be panmictic with those from the South population also suggests the HV has not impacted upon the population structure in Watshamiella. However, the deep divergence inferred between the North and South populations, sampled either side of the BG and StLG, suggests a possible role for these barriers in the divergence history of this NPFW species.

An interesting observation across the divergence times inferred for the small NPFW species is that they all fall within interglacial (or very weak glacial) cycles (with the exception of Sycoscapter short). Assuming the ancestors of F. rubiginosa had already radiated into a transitional species able to grow in dry and/or rocky habitats, this could indicate the warm/wet conditions of the interglacials were unfavourable and/or resulted in increased competition for habitat patches compared with the cool glacial periods. This in turn caused habitat contractions that resulted in the divergence between some of its wasp inhabitants. An alternative scenario is that the signal detected across these species is indicative of range expansions during favourable interglacial conditions. The resultant ancestral population of such
expansions was divided by the unfavourable conditions of the following glacials, leaving behind the signal of divergence seen across these taxa.

### 4.5.1.2.2 Late Pleistocene admixture

A late Pleistocene instantaneous admixture event was inferred in three of the four small NPFW species. The exception to this rule was *Sycoscapter* short, for which a history of continuous migration gave a significantly better fit. The direction of the migration (North to South) in *Sycoscapter* short also differed to that seen in the other three species. *Philotrypesis*, *Watshamiella* and *Eukobelea* all showed South to North admixture events in the late Pleistocene between ~333 kya – 103 kya. Each admixture event occurred during one of the significant interglacials of this period, MIS 5, 7 and 9 (Railsback et al., 2015) for *Watshamiella*, *Philotrypesis* and *Eukobelea*, respectively. The admixture events all occurred prior to noticeable peaks in aridity in Australia (Hocknull et al., 2007). These observations could indicate the events occurred during more favourable conditions at the peaks of these interglacials and that the following severe cooling and consequential drying prevented gene flow from continuing. It has been suggested that the increase in aridity in Australia occurred earlier in the south, compared to the more tropical regions in the northeast coastal areas, evidenced by long term faunal stability (Hocknull et al., 2007). This stability could be indicative of taxa in the south adapting to more arid conditions prior to those in the north. If this is the case, the North populations studied here may have been more vulnerable to the extreme climate changes of the glacial cycles causing drops in population size or possible local extinctions, which were then boosted by individuals from the South during favourable conditions. This is a possible explanation for the South to North direction of admixture inferred for *Philotrypesis*, *Watshamiella* and *Eukobelea*.

### 4.5.1.2.3 Ne – North vs South

The four NPFW species all showed differences in $Ne$ between the North and South populations. However, they did not all agree on which had the larger (or smaller) $Ne$. The North population of *Sycoscapter* short and *Philotrypesis* was larger compared to the South with the opposite being true in *Watshamiella* and *Eukobelea*. As all of these species are associated with the same fig tree, were $Ne$ to reflect habitat patch size in both areas we would expect the ranking of North and South $Nes$ to be consistent across these species. However, my results show that this is not the case, implying that other factors are affecting the $Nes$ in these species. The discrepancy amongst species is perhaps not surprising when considering the continuous
distribution of the host tree *F. rubiginosa*. The habitat and therefore the populations of *F. rubiginosa* are continuous, leading to some doubt over whether (or where) any division into North and South populations of the host fig exists. Although the four NPFW show evidence of significant divergence between the North and South, the boundaries of the populations cannot be determined with the minimal sampling design used here. Also, where populations are sympatric (e.g. in *Philotrypesis*) the tree habitat ranges are shared and so differences in $Ne$ in these areas must reflect other population dynamics or life history traits rather than just the actual habitat area available. For example, an influx of migrants could increase the $Ne$ of a population, while presence of a highly biased sex ratio or a large variance in offspring numbers across the population could decrease $Ne$ (Nielsen and Slatkin, 2013). In three species (*Sycoscapter* short, *Watshamiella* and *Eukobelea*), gene flow is predominantly from the larger $Ne$ population into the smaller $Ne$ population, whereas *Philotrypesis* shows the opposite. It is not clear what difference between these four species might explain this contrasting pattern. The sex ratios in pollinators and NPFW that enter the syconium can be highly female biased, though they are thought to be less so in externally ovipositing NPFW (Wang et al., 2010), which include all four species being considered here. These observations likely suggest several interacting factors are influencing the differences seen in $Ne$ across these populations.

### 4.5.1.3 The wasp inhabitants of *F. rubiginosa*: a community perspective

Overall the results presented here provide a tantalising glimpse into population history variation across the sampled members of the *F. rubiginosa* fig wasp community. Although more comprehensive sampling at the species and population level would be required to obtain a more complete picture of the interactions shaping this community, the reconstruction of the population history between North and South populations obtained for the species sampled here can be related back to the directional vs dispersal driven community assembly hypotheses laid out in section 4.2.6. Whilst neither hypothesis was completely confirmed, some striking patterns are visible across the *F. rubiginosa* community (Figure 4.3). The pollinator and its two parasitoids (*Sycoscapter* long and short) show highly contrasting histories. The divergence without gene flow seen in the pollinator compared to the lack of divergence in *Sycoscapter* long could indicate that the parasitoid is a better long distance disperser compared to its pollinator host (as has been suggested in (Sutton et al., 2016)). However, it is hard to confirm this without appropriate sampling of the pollinator species across the rest of its range, which are also
attacked by *Sycoscapter* long (Segar et al., 2014). If the other pollinator species are not reproductively isolated, then current sampling would miss potential signals of gene flow occurring throughout the range amongst the different members of this species complex. The second parasitoid, *Sycoscapter* short, shows a significantly older divergence time to the pollinator, which does not fit with the directional hypothesis of community assembly (in which parasitoids are predicted to pursue their pollinator hosts through space, and so have more recent divergence times across barriers to gene flow) that was supported for a number of gall wasp parasitoids in the Western Palearctic (Stone et al., 2012). The observed difference in divergence times could result from association of the parasitoid with a different pollinator host in the past such that it could expand its distribution independent of the distribution of the specific pollinator. A similar, more complex explanation is that the population history of *Sycoscapter* short has been influenced by a longer term association with the *Pleistodontes* complex, quite possibly including one or more of the other *Pleistodontes* species not sampled here. Again, more comprehensive sampling of the entire complex may help to shed some light on this aspect of their combined histories.

The histories of both *Sycoscapter* species point to each of them being effective dispersers. The inference of continuous migration in *Sycoscapter* short suggests that gene exchange has been maintained since divergence despite the climate induced habitat changes occurring throughout this time. An even greater dispersal ability could explain the finding that *Sycoscapter* long is a single panmictic population across the sampled range. Sutton et al (2016) used a handful of nuclear microsatellite loci to assess the genetic structure of *Sycoscapter* long throughout its east coast range and found no evidence of differentiation between any of the sampled sites, proposing it may be capable of dispersal distances greater than that of its host. This interpretation has also been applied to a parasitoid wasp (*Hyposoter horticola*) of the Glanville fritillary butterfly in Finland (Couchoux et al., 2016). The authors of this study used 14 microsatellite loci to show the parasitoid maintained a continuous population across the fragmented landscape of its host (Couchoux et al., 2016). It would be interesting to assess the trophic relationship between the two *Sycoscapter* parasitoid species to shed further light on their place within this community.

Assuming the point estimate of the divergence time in *Sycoscapter* short is accurate, the results show that its parasitoid (the hyperparasitoid), *Watshamiella*, has a younger divergence time, which is expected under the directional hypothesis, i.e. a parasitoid will follow its host through space and time (Stone et al., 2012). In the only
other study to assess the genetic structure of hyperparasitoids (i.e. the fourth trophic level), the sampled populations were found to be less differentiated than those of their host parasitoids (Nair et al., 2016). The authors of this study used microsatellite markers to compare the spatial structure across the different trophic levels concluding that genetic structure decreased as trophic level increased. The study area covered was relatively small (50-70 km), compared to the current study (~2000 km), and the overall result was suggested to be due to increased dispersal ability at the higher trophic levels (Nair et al., 2016). It is difficult to propose this explanation for the results here however as the hyperparasitoid, Watshamiella, one of its hosts, Sycoscapter short, and the pollinator, P. imperialis, show structure across populations whereas another of the hyperparasitoids hosts, Sycoscapter long, shows no structure across populations. The addition of temporal information highlights a more complex picture than the spatial structure inferred in the study of Nair et al (2016).

Under the assumption that Sycoscapter short was responding to its host pollinator (i.e. the pollinator had an older divergence time), the results also show that the inquiline, Philotrypesis, has a more recent divergence time compared to the pollinator, which also fits with the fact that Philotrypesis requires the presence of the pollinator to complete its lifecycle.

In contrast, the more ancient divergence time in the small galler, Eukobelea is not compatible with the directional hypothesis: it would be expected that Eukobelea would follow the pollinator as it also requires its presence for completion of its lifecycle. However, interpretation of the population history of Eukobelea also requires more information on the phylogeography of the whole Pleistodontes complex. It is hard to assess concordance of the population history of the large galler, Herodotia, with the community assembly hypotheses, as it may be capable of completing its lifecycle in the absence of the pollinator (Borges, 2015). However, assuming host specificity, it is reliant upon F. rubiginosa, which in turn is reliant upon its pollinator, making Herodotia also reliant on the pollinators in an indirect way.

**4.5.2 Within-guild comparison of population histories**

The population histories inferred for the two Philotrypesis species (i.e. those associated with F. rubiginosa and F. obliqua, respectively) are highly concordant (Figure 4.4). Divergence between the North and South in both is shown to have occurred during the EMPT around the ‘~900 kya event’ (Clark et al., 2006). Philotrypesis (ex F. rubiginosa) is inferred to have diverged during MIS 23 whereas
Philotrypesis (ex F. obliqua) is inferred to have diverged during MIS 21 (Railsback et al., 2015). Both of these stages are representative of warm interglacials that fall during a period that includes the most extreme glacial periods seen in the Pleistocene (Head and Gibbard, 2015a). This could represent a situation where the favourable interglacial conditions allowed a range expansion in these species. During less favourable conditions the populations were divided into two, both of which were maintained throughout subsequent climate cycles. The presence of significant inferred gene flow in both species in the form of an instantaneous admixture event suggests these populations were not, at the time, reproductively isolated. Both admixture events were of a similar magnitude and occurred during the late Pleistocene within subsequent marine isotope stages (MIS 6 and 7 for F. obliqua and F. rubiginosa, respectively) (Head and Gibbard, 2015a). They occurred either side of a proposed peak in aridity in Australia (Hocknull et al., 2007) potentially representing periods during which conditions were favourable enough to facilitate gene flow between these populations. Due to the highly similar ecologies, both currently and historically inferred (Ronsted et al., 2008b), of these two Ficus species, it is easy to imagine them (and congeneric insects associated with each) responding in similar ways to such climate shifts. An interesting discordance in the models inferred for these two species is the contrast between them in North versus South asymmetry in Ne. The South population Ne is larger than the North in Philotrypesis (ex F. obliqua) and vice versa in Philotrypesis (ex F. rubiginosa). It is difficult to attribute either of these differences to patch size as both tree species show relatively continuous distributions (Figures 1.9 and 1.10, Chapter 1) and if anything, the known distribution of F. obliqua seems to suggest more observations of these trees in the North (although the density at each site is unknown) (Dixon et al., 2001). Therefore, as discussed above (section 4.5.1.2.3), it is likely that several interacting factors have shaped the Nes of the populations inferred here. The sampling in both these Philotrypesis species straddles the BG and StLG, suggesting these barriers could have affected the deep divergence inferred here. However, the presence of a large admixture pulse in both also suggests that if they did divide these populations, it was not permanent and these species have been able to exchange genes again at similar times in the past.
4.5.3 Limitations of the composite likelihood inference method

4.5.3.1 Full bSFS vs pairwise composite likelihood inference

The pairwise composite likelihood scheme employed here affords a significant time saving over the full likelihood method used in Chapter 3. The most complex models can be fitted in minutes compared to days (or even weeks) for the full likelihood scheme. This was especially beneficial when conducting a comparative study testing multiple population histories across eight species. It was important however to ensure this time saving did not negatively impact upon the quality of results. For this reason the full dataset used in Chapter 3 was run through all of the equivalent pairwise composite models and the results were compared. The results were highly concordant (Figure 4.1). The same best-fit IM and ADM model were inferred under both schemes, with the composite scheme slightly overestimating the divergence time under both models compared with the full scheme. The biggest noticeable difference was in the ability to significantly separate the best-fit IM and ADM models under each scheme (i.e. the model selection procedure described in the methods). The ADM model was found to fit significantly better over the IM model under the full scheme whereas this could not be confirmed under the composite scheme. The pairwise scheme breaks up sequence blocks into independent observations to greatly simplify the likelihood calculation (Costa and Wilkinson-Herbots, 2017), which results in the loss of the genealogical information gained from linked polymorphic sites. This necessarily reduces the information about past demography. This comparison shows that, despite this trade-off, the pairwise method works well and can estimate parameters under both the IM and ADM model. However, for certain histories, like the young divergence with significant gene flow inferred for *P. nigriventris*, it may not have the power to distinguish between a discrete pulse of admixture and a continuous rate of gene flow.

4.5.3.2 Violations of the composite likelihood assumptions

As with many multi-locus inference methods (and as in Chapter 3), the pairwise composite likelihood method used here assumes no intralocus (i.e. within-block) recombination and free interlocus (i.e. between blocks) recombination. To account for possible violations of these assumptions, I performed a parametric bootstrap for each North/South pair of each species sampled. 95% C.I.s were calculated for each parameter under the best-fit model for each species. The 95% C.I.s are narrow for each parameter, per species model, for all species except *Sycoscapter short*. The divergence time for *Sycoscapter short* has very large C.I.s, indicating that the MLE
for this parameter should be interpreted with caution. The mean estimates obtained from the simulation replicates are all very close to the MLEs used to simulate the datasets, with the exception of the estimate of $T$ in *Sycoscapter short*, suggesting little bias in the estimates. Across all of the species the simulations assumed the same recombination rate ($r$) and mutation rate ($\mu$). If either of these varies across species it would affect how accurate a representation of each the simulations were and also in the case of $\mu$ would affect the scaling of the parameter estimates. The $r$ estimate used here ($2.719 \times 10^{-10}$) was estimated for a pollinating fig wasp (*P. nigriventris*, Chapter 3). The level of inbreeding can affect the effective recombination rate (Charlesworth, 2003) and inbreeding in pollinators is expected to be higher than in NPFW that oviposit externally (Sutton et al., 2016). This is because they are able to lay eggs in more than one fig, increasing the chance of more than one female laying eggs in each fig. The levels of inbreeding in pollinators is dependent upon the number of foundresses in each fig, which makes pollinators of larger figs more likely to show lower levels of inbreeding compared to the inhabitants of smaller figs (J. Cook, personal communication). With these observations in mind, it is possible that the recombination rate used here (that was estimated from a pollinator of a large fig) is an overestimate for the pollinator of *F. rubiginosa* but an underestimate for all of the NPFW. However, the lack of bias seen across simulation replicates for all species, with the exception of *Sycoscapter short*, suggests this is not a bad approximation for the recombination rate for these species.

### 4.5.3.3 Calibrating absolute divergence times

Scaling divergence and admixture times into years requires an estimate of the per site mutation rate ($\mu$) and the number of generations in a year per species. The estimate of $\mu$ ($2.8 \times 10^{-9}$) used here was calculated from *Drosophila melanogaster* (Keightley et al., 2014) as no known estimate is available for fig wasps. It is important to note that the mutation rate could differ from the *Drosophila* one assumed here in fig wasps and potentially across the species sampled, which would affect the estimates of $N_e$, divergence time and time of admixture. However it is somewhat reassuring to note that the only other direct mutation rate estimate in an insect ($2.9 \times 10^{-9}$), which was calculated from *Heliconius melpomene*, is very similar to the one inferred from *Drosophila* (Keightley et al., 2015). The estimates of 4 and 6 generations in a year for the pollinator/gallers and parasitoids/inquilines respectively are based on personal observations (J. Cook, personal communication) and consideration of the differences between the lifecycles across these trophic levels. As discussed in the introduction, the pollinators and gallers lay eggs in the
fig at an earlier developmental stage compared to the parasitoids and inquilines (Segar et al., 2014). The adult wasps across these groups leave the fig at the same time, showing that their pre-adult lifecycles are different lengths (Borges, 2015). The fact that parasitoids and inquilines develop more rapidly than pollinators and gallers, and always have hosts at a suitable stage to attack (because figs and their gall-inducing hosts do not have discrete generations) suggests they could be able to produce more generations in a year. Obtaining an accurate estimate for these species is difficult given their size and lifecycle, meaning these estimates should be treated with an element of caution. A range of estimates (2-6 and 4-8 for pollinators/gallers and parasitoids/inquilines, respectively) were used to assess the differences such estimates would have on the results (Figure 4.5). The biggest impact would be on the divergence time estimates for *Eukobelea*. Assuming 2 rather than 4 generations in a year, the divergence time would be pushed back to the Late Pliocene. It is important to remember though that this uncertainty in these scaling parameters does not affect any of the parameter estimates not scaled by theta or in ‘real time’ (for $\mu$ and generation time respectively), so the best-fit models, presence/absence of gene flow and magnitude of gene flow are all unaffected by the assumptions about the generation time and mutation rate.
**Figure 4.5**: Assessment of the divergence times inferred under different generation time estimates. The thick horizontal black lines represent the times assumed through the results and discussion (using estimates of 4 generations per year for pollinators and gallers and 6 generations per year for parasitoids). The thin horizontal black lines represent each of the times inferred under the range of generation time estimates (2-6 generations per year for pollinators and gallers and 4-8 generations per year for parasitoids).

### 4.5.4 Conclusions

This study has inferred the population histories of several members of the wasp community associated with the fig trees, *F. rubiginosa* and *F. obliqua*. Use of genome-wide data has allowed the estimation of numerous demographic parameters and has provided a level of detail that is thus far unrivalled in studies of Australian fig wasps, or any other east coast Australian taxa. All of the timings inferred relate the histories of these wasps to the fluctuating climate of the Pleistocene, highlighting the likely impact these climate shifts had on these taxa. The within-guild comparison showed remarkable congruence in the histories for two *Philotrypesis* species. Analysis of the *F. rubiginosa* community suggests a complex history in which taxa vary in their sensitivities to climatic events and ecological conditions.
Nevertheless, our observation of young divergence for certain members of the community fits with the extremes of the last glacial maximum while other members diverged much earlier, consistent with the climatic shifts of the Early-Middle Pleistocene transition. As the ability to infer demographic histories improves with the increasing availability of genomic resources and analytical tools, I expect diagnoses of simple co-divergence to become increasingly rare, and be replaced by more complex histories shaped by the interaction of ecological traits and environmental drivers.
Population genomics of a fig wasp community
5 General discussion

5.1 Thesis overview and key findings

The ecologically closed community of fig wasps associated with fig trees is an excellent model system for assessing aspects of community structure and the phylogeographic relationships of co-distributed taxa. I used this system to investigate the phylogeographic histories of trophically-linked fig wasp species from multiple host tree species along the east coast of Australia, a biogeographically diverse latitudinal transect. I took advantage of the availability of high throughput sequencing, and the accompanying advances in laboratory protocols, bioinformatic software and demographic inference methods, to assess these phylogeographic patterns using genome-wide datasets. Below I briefly outline the key findings of this thesis from two perspectives, methodological and phylogeographic.

5.1.1 Methodology

In Chapter 2 I demonstrated that it is possible to obtain genome-wide data from tiny individual wasps, even though many of them did not provide enough DNA or produce appropriate fragment size distributions to meet the quality control recommendations of the genomic library manufacturers. Male pollinating fig wasps are among the smallest insects, implying that similar individual-level approaches should be applicable to the vast majority of insects. New laboratory protocols for preparing genome sequencing libraries from small quantities of DNA are being developed all the time, with more options available now than at the beginning of this study. This will further enable improved genome sequencing of a wide range of small non-model taxa.

The approach used in Chapter 2 provides a further empirical confirmation that population genomic analyses do not need a high coverage reference genome (Hearn et al., 2014). Combining low coverage data across several individuals for each species to improve overall assembly, as was done here, is a cost effective solution that in principle can be applied to any non-model organism. The results in Chapters 3 and 4 highlight how sampling only two individuals per population can be informative for inferring population histories across multiple species. Although such minimal sampling is not new (Li and Durbin, 2011; Lohse et al., 2016), this thesis represents another demonstration of the insights that can be gained from such a sampling approach, a crucial finding for studies of non-model organisms that are perhaps rare and/or difficult to sample. The comparison of the full and composite
likelihood approaches presented in Chapter 4 demonstrates an important practical result in that despite the information lost through the use of the pairwise data summaries of the composite approach, results are highly concordant across both methods. This allows confident use of the composite approach whilst benefitting from the large saving in computational time.

5.1.2 A phylogeographic overview

Despite many phylogeographic studies of east coast Australian taxa (Firestone et al., 1999; James and Moritz, 2000; Pope et al., 2001; Schäuble and Moritz, 2001; Nicholls and Austin, 2005; Brown et al., 2006; Dolman and Moritz, 2006; Edwards and Melville, 2010; Chapple et al., 2011b; Lucky, 2011; MacQueen et al., 2012; Burke et al., 2013; Smissen et al., 2013; Bryant and Fuller, 2014; Hazlitt et al., 2014), the results presented in this thesis are, to my knowledge, the first to employ genome-wide data to infer population histories, and the first to estimate both population divergence times and migration rates between them.

All of the species that showed an inferred history of north-south divergence split during the Pleistocene, a period characterised by known climate fluctuations in the southern hemisphere. Although all divergent events were estimated in the Pleistocene, they spanned the entire epoch, demonstrating high variance across taxa. This suggests that while the same geographic region is associated with population divergence in multiple taxa, such divergence was probably not linked to a single historical event. Although synchronous divergence across sets of co-distributed taxa has been inferred in several studies across the globe (Hickerson et al., 2006; Hickerson and Meyer, 2008; Daza et al., 2010; Chan et al., 2011; Huang et al., 2011; Stone et al., 2012), other studies have inferred a more complex situation of multiple events (Leaché et al., 2007; Plouviez et al., 2009; Voje et al., 2009; Barber and Klicka, 2010; Lawson, 2010) as found here. It has been suggested that community structure is likely too complex to be explained by a simple model of synchronous divergence (Zink et al., 2000; Papadopoulou and Knowles, 2015; Riddle, 2016; Hung et al., 2017) and that differences in species traits (for example, dispersal ability (Joseph et al., 1995; Smith et al., 2014) and ecological specialisation (Joseph et al., 1995; Li et al., 2014; Hung et al., 2017)) contribute to the variance in population structure witnessed across community members. It is therefore possible that the variance in divergence inferred here reflects differences in species traits across trophic levels and how these may have affected each species' response to historical climate change.
Chapters 3 and 4 show that all but one species for which gene flow was inferred experienced gene flow in the same South to North direction. This could result from equal per-individual probabilities of dispersal in both regions but larger populations in the South, or from an asymmetry in probability of dispersal in the two directions (resulting, for example, from directional prevailing winds that could bias dispersal in a particular direction (Calsbeek and Smith, 2003)). South populations have a larger $N_e$ in 3 out of the 4 species to show this South to North gene flow direction. The size of each population could have varied through time however (as demonstrated by the PSMC analysis in Chapter 3), a factor the demographic models that I used here do not incorporate. A shared direction of gene flow was also inferred for a set of co-distributed birds in South America (Oswald et al., 2017). Each species was sampled from dry forest regions separated by the Andes with the suggestion that differing climate conditions between the populations may have affected resource availability in each resulting in the common dispersal pattern (Oswald et al., 2017). Differing climate conditions in the North and South could also explain the pattern seen here, for example, if the North population was repeatedly affected more severely by climatic cycles, causing fluctuations in population size, making admixture from the South more likely. The inference of recent South to North gene flow in the rainforest pollinator, *P. nigriventris*, in Chapter 3, despite the large distance between the two disjunct populations, suggests the potential for major range shifts in these rainforest trees through the late Pleistocene.

The results presented in Chapter 4 show no detectable differentiation between the South and Sydney populations for any *Ficus rubiginosa*-associated wasp species, but significant divergence between the North and each of the South and Sydney populations. This could indicate stronger effects of Pleistocene climate cycles on the distribution of this fig, and its associated wasps, in the northern part of its range. The variation in population histories across trophic levels for wasps associated with a single fig tree, *F. rubiginosa* (Chapter 4), is in contrast to the strikingly similar population histories inferred within the trophic level of the two inquilines (*Philotrypesis*) sampled across two fig trees (*F. rubiginosa* and *F. obliqua*). This result is intriguing and suggestive of a shared response to environmental fluctuations in these two species, resulting in their concordant population histories within a trophic level but across fig tree hosts. The ecologies of these two fig trees are very similar, with *F. obliqua* sometimes described as a 'scaled down' version of *F. rubiginosa* (Darwell, 2013). The morphologies of associated *Philotrypesis* wasps are very similar, although those from the smaller figs of *F. obliqua* are correspondingly smaller insects.
Population genomics of a fig wasp community

(Segar, 2011; Darwell, 2013). This raises the question of whether the similar population histories of the two *Philotrypesis* species result from phylogenetically conserved traits (such as lifespan, reproductive output, or dispersal abilities) likely to influence gene flow (Govindaraju, 1988; Dick et al., 2004; Papadopoulou et al., 2009; Watanabe et al., 2010; Sekar, 2012). It has been noted that their time of attack (i.e. their oviposition sequence compared to the pollinator) matches that of *Philotrypesis* wasps associated with the African fig *F. burtt-davyi*, which suggests the biology of this genus is widely conserved (Compton, 1993; Segar, 2011). Formal analysis would require data on such conserved traits for a wider set of species, analysed in an appropriately phylogenetically-controlled framework (e.g. Whitmee and Orme, 2013; Stevens et al., 2014; Weigelt et al., 2015).

Comparative phylogeographic studies that use genome-wide datasets to infer patterns of population histories across sets of co-distributed taxa are rare. The fact that the species in this study are trophically linked across multiple levels makes it exceptionally rare (I am only aware of one other example to date that assesses such a community using genomic data (Bunnefeld et al., In Preparation)). The first comparative studies across trophic levels focused on phylogenetic concordance, which if observed at high levels was taken as evidence of co-evolution (e.g. Machado et al., 1996; Clark et al., 2000; Spaulding and Von Dohlen, 2001; Clayton et al., 2003). While such reasoning is appropriate for obligate associations (such as termites and their fungal symbionts (Aanen et al., 2002), or aphids and *Buchnera* bacterial symbionts (Baumann et al., 1995)), where the two lineages always co-occur, it does not apply to facultative associations (e.g. corals and their bacterial nitrogen fixing symbionts, (Lema et al., 2012)). In insect host-parasitoid systems, hosts can survive without parasitoids, and parasitoids can often survive without specific hosts, so to understand the extent to which hosts and parasitoids co-evolve, we need to know how often populations of each have interacted through space and time (Hayward and Stone, 2006; Stone et al., 2012; Gebiola et al., 2014). A great deal of focus in comparative phylogeographic studies has been placed on whether population structure is concordant or discordant across co-distributed taxa and discordance has been associated with divergent responses to environmental factors (Voje et al., 2009; Burbrink et al., 2016; Prates et al., 2016; Myers et al., 2017). Such discordance could also be attributed to variation in relevant species traits (e.g. dispersal ability) even though each species has been impacted by the same climate and/or habitat barrier (Papadopoulou and Knowles, 2015; Papadopoulou and Knowles, 2016; Zamudio et al., 2016). This observation highlights the need to
estimate population parameters (e.g. gene flow) under explicit models and to interpret these using knowledge of both individual species traits and historical climate fluctuations (Riddle, 2016; Hung et al., 2017; Oswald et al., 2017). Carrying out this process for sets of non-model taxa, in which prior knowledge of specific traits may be limited, can be difficult, but focusing on a tightly linked multitrophic community that is centred on a common plant structure provides an excellent platform for making meaningful comparisons of phylogeographic histories. For such systems, at least some parameters (such as host plant or host insect distributions) are shared across taxa. This contrasts with the majority of comparative phylogeographic analyses, which compare patterns across potentially ad hoc collections of co-distributed taxa that differ widely in many traits likely to influence population history (Lapointe and Rissler, 2005; Feldman and Spicer, 2006; Leaché et al., 2007; Hickerson and Meyer, 2008; Jones and Kennedy, 2008; Chan et al., 2011; Page and Hughes, 2014; Burbrink et al., 2016).

5.2 Study limitations, and what I’d do differently (hindsight is a wonderful thing)

5.2.1 Sampling

The taxa in Chapter 4 were chosen to maximise the number of species sampled given the resources available. However, sampling and sequencing a smaller number of taxa in more depth might have yielded more resolved population histories. Ultimately the sampling approach involved a trade-off between high resolution analysis of single-species histories on one hand and (the choice taken here) the ability to compare across species on the other. However, the presence of so many cryptic taxa within the morpho-species sampled makes community level inferences difficult, as the sampled species represent only a portion of the lineages involved in the histories of the affected species. For example, complete sampling of all five proposed Pleistodontes imperialis species could have provided the opportunity to understand both their phylogeographic histories and the deeper phylogenetic relationships between these sister species. Sampling for such an approach would require careful consideration given the complex distributions of the five putative species and the differences in methodological requirements for a phylogenomic analyses (e.g. complete sampling of the species complex plus an appropriate outgroup, a more targeted sequencing approach to obtain orthologous loci across all samples). The limited time available and the small proportion of the whole geographic area covered for sample collection in this study were limiting factors in
obtaining the appropriate samples, especially as one of the species appears to be extremely rare and has only ever been sampled at low levels (Darwell, 2013).

The minimal individual-level sampling employed here, whilst being beneficial for sampling rare species and for maximising the number of species sampled within the available resources, does restrict the methods available for analysis. For example, any method that is based upon the site frequency spectrum (SFS) (e.g. dadi (Gutenkunst et al., 2009) and Fastsimcoal2 (Excoffier et al., 2013)) requires a larger number of individuals sampled per population to be able to calculate allele frequencies. Sampling of more individuals would have enabled a potentially informative comparison over a larger range of methods. However, given the resource limitations of this study, this alternative approach would have restricted sampling to a smaller number of species/populations. Larger sample sizes would also not have been possible for some of the species studied, e.g. the small gallers of the genus *Eukobelea* for which a total of only 5 samples were available. For rare species, inference methods that make the most of minimal sampling allow important insights to be gained. This advantage could also be extremely relevant to the study of endangered species (as suggested in Nazareno et al (2017)) and to studies that involve ancient DNA, e.g. establishing the population history shared between modern humans and Neanderthals (Lohse and Frantz, 2014).

5.2.2 Inclusion of selection into demographic models

As discussed in section 1.2, analyses of genetic diversity within and among genomes lie at the heart of methods for detecting both selection and inferring population demographic history. However, the genomic data requirements differ between inference of selection and history. Genome scan methods for detection of selection (introduced in Chapter 1) require sampling of a large number of individuals to allow areas of high Fst (or another summary statistic) to be tested for statistical significance. They also require more fully assembled genomes (i.e. less fragmented) to allow for a continual sliding window analysis along each chromosome (these are often done at a resolution of 10-50 kb, requiring much longer contigs than those generated for this study e.g. only 0.2% of the contigs generated in the highest quality assembly of this study (*P. nigriventris*) were over 50 kb in length). Finally, identification of candidate loci under selection requires an annotated reference genome to be available for the species under study, so that the phenotypic significance of any outlier loci and their potential role in adaptive divergence between populations can be inferred. Given the aim of this study, the system used
and the resources available, it was not possible to extend the analysis to address questions relating to selection. However, this could form an interesting future direction for the fig wasp system along the latitudinal transect of eastern Australia.

5.2.3 Bioinformatics

The bioinformatic approaches chosen here were those thought to be the most appropriate for each task. The number of options available for all areas of bioinformatic processing increases almost daily, given the rise in interest and availability of high throughput sequencing data, which sometimes makes choosing the best option overwhelming (Smith, 2015; Duck et al., 2016). If time and computing power were limitless, a comparison of several approaches would be a comprehensive way to assess the best option, weighing up the technical differences and potential biases between each (e.g. Liu et al., 2013; Hwang et al., 2015).

Currently the time required to run each stage of a bioinformatic processing pipeline can be a problem even once each piece of software is chosen. This is because many processes take a long time to run (e.g. the SPAdes assemblies used in this study took several weeks to run, even when parallelised over multiple threads on a high performance server) and they all require the user to specify values for the input parameters (e.g. quality filter thresholds) so cannot be fully automated. Ideally the effects on the data of varying these values should be investigated (e.g. Schilling et al., 2014), however in this study I did not have time to assess the possible impacts of such choices. Values were chosen based on prior experiences of other researchers and/or on the recommendations of the method’s authors.

The assumption is that over the range I used, block length should not affect the results. However, the longer the blocks become, the more likely they are to have been affected by past recombination events, which would therefore place a limit on suitable block lengths (Hey and Nielsen, 2004; Wang and Hey, 2010; Lohse et al., 2011; Lohse et al., 2016). As the likelihood calculations assume no intra-block recombination, those that fail the 4-gamete test are filtered out of the analysis, which would result in a reduction of the amount of data and hence the power available for inference. Unfortunately, I was also unable to carry out this analysis due to time constraints.

The contaminant filtering steps carried out on the data in Chapter 2 used very broad classifications (i.e. at the phylum level) in order to remove reads potentially not originating from a fig wasp. A more comprehensive contaminant screen could have been carried out to assess the bacterial/fungal communities associated with
different fig wasp species. This is a potentially interesting add-on project for many high throughput sequencing studies as it has been shown that many assemblies harbour microbial sequences that tend to be ignored (Niu et al., 2015). The high likelihood of sequencing untargeted microbes was taken advantage of by Niu et al (2015) when they carried out a screen of the newly sequenced genome of the fig wasp, *Ceratosolen solmsi*, and examined the fungal community that was present within the individuals sequenced. The addition of further individuals to the current study could allow a geographic and multistrophic assessment of the incidence and prevalence of inherited bacterial symbionts (such as *Wolbachia*, *Cardinium*, *Spiroplasma*, *Flavobacteria* and *Rickettsia*), which are widespread in Hymenoptera and have a range of phenotypic impacts on infected hosts (Clark, 1982; Zchori-Fein and Perlman, 2004; Hurst and Jiggins, 2005; Duron et al., 2008; Weinert et al., 2009).

An approach that could have been employed to carry out a phylogenetic analysis of the sampled taxa, as opposed to the analysis of average genetic distance using genome-wide SNPs done in Chapter 2, would have been to use the conserved regions identified by the CEGMA/BUSCO analyses of the reference assemblies (e.g. Luo et al., 2015; Husnik and McCutcheon, 2016). Identifying the regions common to all individuals across the multiple assemblies would have made it possible to use other sampled species as outgroups in a phylogenetic analysis without the requirement of re-aligning the reads to the appropriate reference. Such an approach would be analogous to ‘gene capture’ studies that use specific DNA target baits to capture a specific set of genes in multiple taxa (e.g. Nicholls et al., 2015). However, it is not clear whether there would be enough phylogenetic signal in such data to assess relationships between closely related individuals using this set of highly conserved genes.

5.2.4 All models are wrong but some are useful

‘All models are wrong but some are useful’ (Box and Draper, 1987) is an important message to remember when interpreting results based on models that by definition represent over-simplified versions of real world scenarios (Hickerson, 2014). The pairwise divergence models explored in Chapters 3 and 4 were designed to gain insight into divergence and gene flow, processes that are known to affect populations through time. Naturally occurring populations are likely to exhibit more complex structure than the simple pairwise comparisons that form the basis of the modelling framework employed here. However, *P. nigriventris* (Chapter 3) is the only known pollinator of the rainforest fig tree, *F. watkinsiana*, which is found in two
distinct populations with no significant patches of trees in between. In this case a two-population model seems highly appropriate and a more complex model including more populations may not add anything to the results in Chapter 3. However, further sampling from within each of these distinct populations would enable tests to detect the presence of any sub-structure within them. One way of testing this, as used in Hearn et al (2014), is to explore whether the inferred population history between two regions depends on which individuals from across each region are included in the analysis. If the inferred history does not change with the different individuals used, then it is unlikely that significant substructure exists within each region. Despite the results in Chapter 4 (F. rubiginosa-associated species) suggesting the South and Sydney regions can effectively be treated as a single population, the boundaries of the North/South populations are not clear cut due to the continuous distribution of the host tree. Therefore the two population model does not necessarily capture the full picture in these species, and wider geographic sampling and appropriate methods that could deal with increased sample numbers would be required to investigate this further (e.g. ABC or SFS approaches).

A best fit model only indicates that, out of the set of models fitted to the data, this is the one that most closely matches the signal in the data. For this reason, the choice of the set of models tested is highly relevant to the resulting population history inferred (Nielsen and Beaumont, 2009). Model selection should be based upon a priori knowledge of the taxa under study and therefore the appropriateness of the model set selected will likely depend on the amount of previous research associated with those taxa (Carstens et al., 2013; Pelletier and Carstens, 2014; Thomé and Carstens, 2016). This is obviously more of a problem in non-model taxa in which a phylogeographic study using molecular data may be one of the first investigations involving that particular species (Thomé and Carstens, 2016). This situation contrasts with that for humans where many detailed studies have been conducted for decades and data gathered across many disciplines to aid the choice of the most appropriate models. The amount of data available to those working on human demography is unrivalled, for example the large sequencing projects that aim to fill databases with whole human genomes sampled from around the world (International Human Genome Sequencing Consortium, 2001; International SNP Map Working Group, 2001; International HapMap Consortium, 2005; 1000 Genomes Project Consortium, 2015). These provide high quality resources for a multitude of large-scale studies covering various aspects of human demography (Marth et al., 2004; Williamson et al., 2005; Bryc et al., 2010; Wall et al., 2011; Lachance et al., 2012;
Fu et al., 2013; Gazave et al., 2014; Raghavan et al., 2015; Omrak et al., 2016; Pugach et al., 2016). The data resources are not purely sequence based however, they are also derived from extensive anthropological and archaeological evidence that contribute important insight into human population history (e.g. Bräuer, 2008; Hublin et al., 2017). With this wealth of knowledge and usable data comes the ability to fit models far greater in complexity than the ones used in this thesis (Fagundes et al., 2007; Gronau et al., 2011; Yang et al., 2012). Such complex parameter-rich models are able to discriminate among more realistic scenarios that are likely to include more nuanced representations of how populations change over time, and could contain, for example, multiple admixture events or population size changes through time (e.g. Blum and Jakobsson, 2011; Lukic and Hey, 2012; Chen et al., 2015).

Despite the ability to model more complex histories with more data, there will always be a limit to what can be inferred given the data available. To be able to meaningfully fit a model to a dataset there must be sufficient signal present in those data (Stocks et al., 2014). The most effective way of addressing this issue is through the use of simulations. This can be an extremely useful planning tool (Hoban et al., 2012). For example, I could use simulations based on observed data to assess how the signal present under different scenarios would change with increased sampling of individuals and/or populations and sample or sequence accordingly (e.g. Robinson et al., 2014).

Simulated datasets would also be the best approach for assessing the effects of using incorrect estimates of descriptive parameters such as mutation rate and recombination rate (Cornuet et al., 2010; Orlando et al., 2013; Field et al., 2016; Seoighe and Scally, 2017). Both the analyses in Chapters 3 and 4 use the same mutation and recombination rates across all of the species studied due to a lack of species-specific values. Simulations provide the only way (given current information) of assessing how cross-species variation in these parameters might influence my results. A further point that simulations could help to address is whether the size of each dataset (in terms of number and size of sequence blocks), which varies across species, affects the population history inferred (Bunnefeld et al., 2015). Unfortunately due to the time consuming and computationally demanding nature of generating such simulations I was unable to carry out such assurance checks on the data presented here.
The calibrations used in Chapters 3 and 4 to convert divergence and admixture times from generations into years assume knowledge of the mutation rate and generation time for each species. The mutation rate used was estimated in Drosophila melanogaster (Keightley et al., 2014) and the generation times used were estimated from personal observations and knowledge of the life cycles of the different trophic levels of these wasps (J. Cook, personal communication). Given likely inaccuracy of these estimates, divergence and admixture times in years should be treated cautiously. It is important to note that such calibration uncertainty applies to any demographic model and is not specific to the analytical approaches chosen in this thesis. It is also important to reiterate that as long as values of these parameters (while unknown) are similar across species, then comparisons of relative ages of events across species are still valid. The difficulty only lies in assessing the absolute ages of events, a common problem in non-model taxa (Stone et al., 2012; Hearn, 2014; Chen et al., 2016; Oswald et al., 2017; Satler and Carstens, 2017). Divergence times in generations provide an absolute upper limit to population divergence times, as all of the fig wasp species studied exhibit more than one generation in a year. If these upper limits are assumed to be the divergence times for the species in Chapter 4, it makes no change to the order of divergence times except that divergence in Sycoscapter short would be older than in Eukobelea (though very wide confidence limits for Sycoscapter short mean that estimates in these two species overlap whether numbers of generations per year are taken into consideration or not).

5.3 Future directions
The comparative approach adopted in Chapter 4 is relevant to other food web focused communities. For example, a very similar approach is being applied to a set of western Palearctic oak gall wasps (Hymenoptera, Cynipidae) and their associated chalcid parasitoids (Bunnefeld et al., In Preparation). The gallwasp study uses low coverage genome-wide datasets to assess the phylogeographic histories of multiple trophic levels of gall wasps, sampled along a longitudinal transect spanning the Western Palaearctic from Spain to Iran. This area is known to contain multiple refugia for oaks and associated insect species during the Pleistocene glacial cycles (Petit et al., 2002; Nicholls et al., 2010b; Bihari et al., 2011; Lohse et al., 2012; Stone et al., 2012). The genetic consequences of longitudinal range expansion and gene flow between refugia are evident in the genomes of extant individuals, and this study aims to assess whether gallwasps and their parasitoid enemies share similar population histories. As well as being applied to other enclosed communities of
herbivorous insects (e.g. yucca moths, (Althoff, 2008), aphids, (Müller et al., 1999), leafminers, (Memmott et al., 1994; Rott and Godfray, 2000)), the approach used in Chapter 3 could be appropriate for investigating the phylogeographic histories of other taxa inhabiting the same rainforest populations as P. nigriventris in Queensland, Australia. Establishing the histories of a wide range of taxa that share these distinct rainforest populations will provide a more detailed picture of the processes that have led to this split distribution of rainforest taxa and may highlight the species that are at risk if these habitats continue to diminish under the continued drying of the continent (see, Moritz and Faith, 1998; Smith et al., 2000; Carnaval et al., 2009; Igea et al., 2013; Dussex et al., 2014, for examples of phylogeographic studies that make suggestions for conservation priorities).

As alluded to above, the difficulties associated with sampling multi trophic communities that include cryptic species have left us with only a partial picture of the phylogeographical histories of these fig wasp genera along the east coast of Australia. The pollinator of F. rubiginosa, P. imperialis, and its associated parasitoids of the genera Sycoscapter, Philotrypesis and Watshamiella, would make excellent candidates for increased sampling and further analyses to investigate both the phylogeographic and phylogenetic relationships between sister species. Combining such approaches would help shine light on how these species came to occupy their largely sympatric distributions and on the deeper evolutionary relationships present between them.

A further interesting use of increased sampling across this latitudinal transect would be to choose a couple of species (potentially two pollinators from different trees or one pollinator and its parasitoid from the same tree) with which to collect further individuals from both the North and South populations and sequence these to a higher depth, increasing the resolution across larger sections of the genome and allowing a genome scan to look for areas under selection to be carried out. A large latitudinal transect such as the east coast of Australia is an excellent candidate to carry out such a study as there is a strong temperature gradient present along its length, with habitats ranging from cool and dry forests in the South to warm and wet rainforests in the North (Hoffmann and Weeks, 2007). Several studies have looked at different genes associated with a range of traits in Drosophila species along this transect (Hoffmann and Weeks, 2007; Rako et al., 2007; Liefting et al., 2009). They suggest that the changing climatic conditions are exerting different selection pressures which results in phenotypic differences found amongst these predominantly continual populations (Hoffmann and Weeks, 2007; Rako et al., 2007;
Liefting et al., 2009). Given the significant signal of population divergence inferred here for the two distinct \textit{P. nigriventris} populations it would be interesting to ask whether there are any strong signals of gene divergence present that have contributed to the isolation of these populations.

An interesting accompaniment to further sampling of the fig wasp communities would be to extend the sampling to other trophically-linked groups – particularly the trees themselves. For example, generation of phylogeographic population models for the fig tree species would allow us to address such questions as (i) does \textit{F. watkinsiana} show a concordant history of divergence with its pollinator? And (ii) does the population structure of \textit{F. rubiginosa} help explain the North/South divergences inferred in six out of seven of its sampled fig wasp species? The answers to these questions could help establish whether the splits detected across these fig wasp species are likely directly related to their tree hosts’ responses to climate fluctuations or whether they are more likely attributed to other factors.

Different approaches would need to be established to infer such histories in fig trees however. Plants often have very large genomes, often a result of polyploidy, which creates additional sequencing challenges. A genomic reduction technique, as mentioned in Chapter 2 (e.g. Nicholls et al (2015) for trees in the genus \textit{Inga}), could be an appropriate option to overcome this. Also, the highly discordant ecologies of fig trees compared to their associated wasp populations is an important consideration when designing experiments to infer their population histories (e.g. large generation times, low population densities, dispersal abilities through both pollen transfer and seed transport by frugivores).

Figs are home to a diverse range of organisms as well as their wasp inhabitants. Several species of nematode are specific to particular pollinating fig wasp hosts, and rely on them for transport between figs (Herre, 1993; Jauharlina et al., 2012). The nematodes are either phytophagous or parasitic, killing their pollinating fig wasp host directly following transport (Krishnan et al., 2010). They have been shown to affect offspring numbers and the dispersal ability of female pollinators (Herre et al., 2008). Their highly species-specific interaction with fig wasps is suggested to be representative of an ancient association between these three trophic levels (Herre, 1993) leading to the question of what, if any, role nematodes play in maintaining the fig/fig wasp mutualism. Distinct and diverse fungal communities have also been shown to be associated with figs and fig wasps (Martinson et al., 2012; Niu et al., 2015) adding yet another layer of complexity to these enclosed microcosms. These aspects of the fig community are relatively understudied compared to their
pollinating wasps but are important components to take into account when considering the community as a whole. Establishing the histories of and relationships between these additional components of the fig/fig wasp system will add to our knowledge of this ancient and fascinating system.

The focus of this proposed future work has been on the Australian fig wasp community studied in this thesis however it is important to note its wider relevance to the global study of this fig centred system (Azuma et al., 2010; McLeish et al., 2010; Warren et al., 2010; Wei et al., 2014; Heer et al., 2015). It will be important to link the patterns of fig wasp diversification and community structure established in Australian species to those of analogous systems across the entire natural range of *Ficus* and to possibly extend these studies to incorporate its expanding non-native range.

### 5.4 Concluding remarks

This thesis demonstrates how low coverage genome-wide data can be successfully applied to study the demographic histories of co-distributed taxa spanning multiple trophic levels. The results suggest that the Pleistocene epoch was extremely influential for fig wasp communities in eastern Australia despite its characteristic glacial/interglacial cycles being less severe in the southern hemisphere compared to the north. Sampling across host fig trees highlights that despite their contrasting ecologies a divergence history with gene flow was the most common model inferred for their wasp inhabitants. A high level of variance was detected across trophic levels associated with the same fig tree host but striking concordance among species that occupy the same trophic level across host tree species. This suggests that the trophic level is influential in species’ responses to environmental fluctuations. This thesis offers a first glimpse of the population history of Australian fig wasp communities using genomic scale data. It has highlighted interesting areas for future work, all of which will benefit from the genomic resources and phylogeographic insights generated here.
Bibliography


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Compton, S. G. (1993). One Way to Be a Fig. *African Entomology*, 1, 151-158.


Conchou, L., Ciminera, M., Hossaert-McKey, M. & Kjellberg, F. (2014). The Non-Pollinating Fig Wasps Associated with *Ficus guianensis*: Community Structure and Impact of the Large Species on the Fig/Pollinator Mutualism. *Acta Oecologica*, 57, 28-37.


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Bibliography

241
Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


248 Bibliography


Ronsted, N., Weiblen, G. D., Clement, W. L., Zerega, N. J. C. & Savolainen, V. (2008a). Reconstructing the Phylogeny of Figs (Ficus, Moraceae) to Reveal the History of the Fig Pollination Mutualism. v. 45.


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community


Population genomics of a fig wasp community
Appendix A – Bioanalyser traces
Population genomics of a fig wasp community
Appendix A: Bioanalyser traces for each of the sequenced individuals.
Population genomics of a fig wasp community
## Appendix B – Library pooling for sequencing

<table>
<thead>
<tr>
<th>Species</th>
<th>Sample Library</th>
<th>Index 1</th>
<th>Index 2</th>
<th>ng/ul (average from 2 qubits)</th>
<th>ng (concentration X volume)*</th>
<th>Average bp (from BioA)</th>
<th>nM (concentration X 1515)/average bp</th>
<th>Volume of library for 25ul of 2nM</th>
<th>Volume of buffer (Tris + Tween)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleistodontes</td>
<td>FON_403_POL</td>
<td>N507</td>
<td>N704</td>
<td>1.31</td>
<td>34.06</td>
<td>571</td>
<td>3.48</td>
<td>14.4</td>
<td>10.6</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FON_398_POL</td>
<td>N507</td>
<td>N703</td>
<td>2.76</td>
<td>71.76</td>
<td>427</td>
<td>9.79</td>
<td>5.1</td>
<td>19.9</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FOS_183_POL1</td>
<td>N501</td>
<td>N705</td>
<td>1.11</td>
<td>29.97</td>
<td>775</td>
<td>2.17</td>
<td>23.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FOS_187_POL</td>
<td>N507</td>
<td>N706</td>
<td>1.9</td>
<td>49.40</td>
<td>668</td>
<td>4.31</td>
<td>11.6</td>
<td>13.4</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRN_364_POL</td>
<td>N501</td>
<td>N703</td>
<td>2.66</td>
<td>71.82</td>
<td>471</td>
<td>8.56</td>
<td>5.8</td>
<td>19.2</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRN_J7_POL</td>
<td>N501</td>
<td>N704</td>
<td>2.65</td>
<td>71.55</td>
<td>455</td>
<td>8.82</td>
<td>5.7</td>
<td>19.3</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRS_234_POL</td>
<td>N501</td>
<td>N701</td>
<td>2.28</td>
<td>61.56</td>
<td>522</td>
<td>6.62</td>
<td>7.6</td>
<td>17.4</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRS_511_POL</td>
<td>N501</td>
<td>N702</td>
<td>3.34</td>
<td>90.18</td>
<td>490</td>
<td>10.33</td>
<td>4.8</td>
<td>20.2</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FON_339_SYC</td>
<td>N503</td>
<td>N701</td>
<td>3.78</td>
<td>102.06</td>
<td>815</td>
<td>7.03</td>
<td>7.1</td>
<td>17.9</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FON_TrinBeach_ SYC</td>
<td>N503</td>
<td>N704</td>
<td>3.91</td>
<td>105.57</td>
<td>788</td>
<td>7.52</td>
<td>6.7</td>
<td>18.3</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FOS_186_SYC</td>
<td>N503</td>
<td>N705</td>
<td>5.35</td>
<td>144.45</td>
<td>789</td>
<td>10.27</td>
<td>4.9</td>
<td>20.1</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FOS_199_SYC</td>
<td>N503</td>
<td>N706</td>
<td>5.17</td>
<td>139.59</td>
<td>805</td>
<td>9.73</td>
<td>5.1</td>
<td>19.9</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRN_350_SYC</td>
<td>N502</td>
<td>N704</td>
<td>1.15</td>
<td>31.05</td>
<td>590</td>
<td>2.95</td>
<td>16.9</td>
<td>8.1</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRN_274_SYC1</td>
<td>N508</td>
<td>N703</td>
<td>0.965</td>
<td>25.09</td>
<td>665</td>
<td>2.20</td>
<td>22.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Species</td>
<td>Sample Library</td>
<td>Index 1</td>
<td>Index 2</td>
<td>ng/ul (average from 2 qubits)</td>
<td>ng (concentration X volume)*</td>
<td>Average bp (from BioA)</td>
<td>nM (concentration X 1515)/average bp</td>
<td>Volume of library for 25ul of 2nM</td>
<td>Volume of buffer (Tris + Tween)</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td>---------</td>
<td>---------</td>
<td>------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRS_32_SYC</td>
<td>N502</td>
<td>N705</td>
<td>2.54</td>
<td>68.58</td>
<td>658</td>
<td>5.85</td>
<td>8.5</td>
<td>16.5</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRS_397_SYC</td>
<td>N508</td>
<td>N701</td>
<td>2.42</td>
<td>62.92</td>
<td>499</td>
<td>7.35</td>
<td>6.8</td>
<td>18.2</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FON_340_PHIL</td>
<td>N504</td>
<td>N701</td>
<td>3.62</td>
<td>97.74</td>
<td>553</td>
<td>9.92</td>
<td>5.0</td>
<td>20.0</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FON_PDoug_PHIL</td>
<td>N504</td>
<td>N702</td>
<td>4.04</td>
<td>109.08</td>
<td>739</td>
<td>8.28</td>
<td>6.0</td>
<td>19.0</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FOS_186_PHIL</td>
<td>N504</td>
<td>N703</td>
<td>0.779</td>
<td>21.03</td>
<td>704</td>
<td>1.68</td>
<td>29.8</td>
<td>-4.8</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FOS_188_PHIL1</td>
<td>N508</td>
<td>N705</td>
<td>1.645</td>
<td>42.77</td>
<td>614</td>
<td>4.06</td>
<td>12.3</td>
<td>12.7</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRN_310_PHIL</td>
<td>N504</td>
<td>N705</td>
<td>1.99</td>
<td>53.73</td>
<td>747</td>
<td>4.04</td>
<td>12.4</td>
<td>12.6</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRN_407_PHIL</td>
<td>N508</td>
<td>N706</td>
<td>2.04</td>
<td>53.04</td>
<td>586</td>
<td>5.27</td>
<td>9.5</td>
<td>15.5</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRS_247_PHIL</td>
<td>N505</td>
<td>N704</td>
<td>2.88</td>
<td>77.76</td>
<td>642</td>
<td>6.80</td>
<td>7.4</td>
<td>17.6</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRS_110_PHIL</td>
<td>N505</td>
<td>N703</td>
<td>2.63</td>
<td>71.01</td>
<td>714</td>
<td>5.58</td>
<td>9.0</td>
<td>16.0</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRS_511_PHIL</td>
<td>N505</td>
<td>N705</td>
<td>4.05</td>
<td>109.35</td>
<td>635</td>
<td>9.66</td>
<td>5.2</td>
<td>19.8</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRS_40_PHIL</td>
<td>N501</td>
<td>N707</td>
<td>0.855</td>
<td>22.23</td>
<td>533</td>
<td>2.43</td>
<td>20.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRN_299_WAT</td>
<td>N505</td>
<td>N706</td>
<td>4.12</td>
<td>111.24</td>
<td>536</td>
<td>11.65</td>
<td>4.3</td>
<td>20.7</td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRN_337_WAT</td>
<td>N501</td>
<td>N708</td>
<td>1.34</td>
<td>34.84</td>
<td>483</td>
<td>4.20</td>
<td>11.9</td>
<td>13.1</td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRS_239_WAT</td>
<td>N506</td>
<td>N703</td>
<td>2.38</td>
<td>64.26</td>
<td>505</td>
<td>7.14</td>
<td>7.0</td>
<td>18.0</td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRS_514_WAT</td>
<td>N508</td>
<td>N702</td>
<td>2.06</td>
<td>55.62</td>
<td>474</td>
<td>6.58</td>
<td>7.6</td>
<td>17.4</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRN_261_HERO</td>
<td>N506</td>
<td>N705</td>
<td>3.6</td>
<td>97.20</td>
<td>539</td>
<td>10.12</td>
<td>4.9</td>
<td>20.1</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRN_349_HERO</td>
<td>N506</td>
<td>N706</td>
<td>2.31</td>
<td>62.37</td>
<td>452</td>
<td>7.74</td>
<td>6.5</td>
<td>18.5</td>
</tr>
<tr>
<td>Species</td>
<td>Sample Library</td>
<td>Index 1</td>
<td>Index 2</td>
<td>ng/ul (average from 2 qubits)</td>
<td>ng (concentration X volume)*</td>
<td>Average bp (from BioA)</td>
<td>nM (concentration X 1515)/average bp</td>
<td>Volume of library for 25ul of 2nM</td>
<td>Volume of buffer (Tris + Tween)</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------</td>
<td>---------</td>
<td>---------</td>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRN_372_HERO</td>
<td>N501</td>
<td>N709</td>
<td>0.764</td>
<td>20.63</td>
<td>515</td>
<td>2.25</td>
<td>22.2</td>
<td>2.8</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRS_110_HERO</td>
<td>N507</td>
<td>N701</td>
<td>0.972</td>
<td>26.24</td>
<td>499</td>
<td>2.95</td>
<td>16.9</td>
<td>8.1</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRS_239_HERO</td>
<td>N507</td>
<td>N702</td>
<td>1.35</td>
<td>36.45</td>
<td>476</td>
<td>4.30</td>
<td>11.6</td>
<td>13.4</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRS_517_HERO</td>
<td>N501</td>
<td>N710</td>
<td>1.635</td>
<td>44.15</td>
<td>485</td>
<td>5.11</td>
<td>9.8</td>
<td>15.2</td>
</tr>
<tr>
<td>Eukobelea</td>
<td>FRN_310_EUKO</td>
<td>N501</td>
<td>N711</td>
<td>1.24</td>
<td>33.48</td>
<td>661</td>
<td>2.84</td>
<td>17.6</td>
<td>7.4</td>
</tr>
<tr>
<td>Eukobelea</td>
<td>FRN_406_EUKO</td>
<td>N501</td>
<td>N712</td>
<td>0.505</td>
<td>13.64</td>
<td>532</td>
<td>1.44</td>
<td>34.8</td>
<td>-9.8</td>
</tr>
<tr>
<td>Eukobelea</td>
<td>FRS_41_EUKO</td>
<td>N502</td>
<td>N707</td>
<td>1.175</td>
<td>31.73</td>
<td>499</td>
<td>3.57</td>
<td>14.0</td>
<td>11.0</td>
</tr>
<tr>
<td>Eukobelea</td>
<td>FRS_239_EUKO</td>
<td>N502</td>
<td>N708</td>
<td>2.015</td>
<td>54.41</td>
<td>423</td>
<td>7.22</td>
<td>6.9</td>
<td>18.1</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRSYD_1065_PO</td>
<td>N502</td>
<td>N709</td>
<td>1.42</td>
<td>38.34</td>
<td>459</td>
<td>4.69</td>
<td>10.7</td>
<td>14.3</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FRSYD_2884_PO</td>
<td>N505</td>
<td>N708</td>
<td>2.4</td>
<td>64.8</td>
<td>500</td>
<td>7.27</td>
<td>6.9</td>
<td>18.1</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRSYD_89_SYC</td>
<td>N502</td>
<td>N711</td>
<td>1.805</td>
<td>48.735</td>
<td>537</td>
<td>5.09</td>
<td>9.8</td>
<td>15.2</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRSYD_2873_SY</td>
<td>N502</td>
<td>N712</td>
<td>0.762</td>
<td>20.574</td>
<td>707</td>
<td>1.63</td>
<td>30.6</td>
<td>-5.6</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRN_332_SYC</td>
<td>N503</td>
<td>N707</td>
<td>0.901</td>
<td>24.327</td>
<td>552</td>
<td>2.47</td>
<td>20.2</td>
<td>4.8</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRS_238_SYC</td>
<td>N503</td>
<td>N708</td>
<td>1.345</td>
<td>36.315</td>
<td>512</td>
<td>3.98</td>
<td>12.6</td>
<td>12.4</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRSYD_1941_SY</td>
<td>N503</td>
<td>N709</td>
<td>1.37</td>
<td>36.99</td>
<td>450</td>
<td>4.61</td>
<td>10.8</td>
<td>14.2</td>
</tr>
<tr>
<td>Sycoscapter</td>
<td>FRS_508_SYC</td>
<td>N503</td>
<td>N710</td>
<td>0.861</td>
<td>23.247</td>
<td>437</td>
<td>2.98</td>
<td>16.8</td>
<td>8.2</td>
</tr>
<tr>
<td>Species</td>
<td>Sample Library</td>
<td>Index 1</td>
<td>Index 2</td>
<td>ng/ul (average from 2 qubits)</td>
<td>ng (concentration X volume)*</td>
<td>Average bp (from BioA)</td>
<td>nM (concentration X 1515)/average bp</td>
<td>Volume of library for 25ul of 2nM</td>
<td>Volume of buffer (Tris + Tween)</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------</td>
<td>---------</td>
<td>---------</td>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td>------------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRSYD_1822_PHIL</td>
<td>N503</td>
<td>N712</td>
<td>0.391</td>
<td>10.557</td>
<td>611</td>
<td>0.97</td>
<td>51.6</td>
<td>-26.6</td>
</tr>
<tr>
<td>Philotrypesis</td>
<td>FRSYD_165_PHIL</td>
<td>N505</td>
<td>N710</td>
<td>1.09</td>
<td>29.43</td>
<td>605</td>
<td>2.73</td>
<td>18.3</td>
<td>6.7</td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRSYD_567_WAT</td>
<td>N505</td>
<td>N711</td>
<td>1.875</td>
<td>50.625</td>
<td>537</td>
<td>5.29</td>
<td>9.5</td>
<td>15.5</td>
</tr>
<tr>
<td>Watshamiella</td>
<td>FRSYD_1905_WAT</td>
<td>N505</td>
<td>N712</td>
<td>1.785</td>
<td>48.195</td>
<td>478</td>
<td>5.66</td>
<td>8.8</td>
<td>16.2</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRSYD_381_HERO</td>
<td>N504</td>
<td>N709</td>
<td>1.995</td>
<td>53.865</td>
<td>509</td>
<td>5.94</td>
<td>8.4</td>
<td>16.6</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRSYD_293_HERO</td>
<td>N504</td>
<td>N710</td>
<td>2.15</td>
<td>58.05</td>
<td>492</td>
<td>6.62</td>
<td>7.6</td>
<td>17.4</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRSYD_1964_HERO</td>
<td>N504</td>
<td>N711</td>
<td>1.37</td>
<td>36.99</td>
<td>512</td>
<td>4.05</td>
<td>12.3</td>
<td>12.7</td>
</tr>
<tr>
<td>Herodotia</td>
<td>FRS_244_HERO</td>
<td>N504</td>
<td>N712</td>
<td>2.57</td>
<td>69.39</td>
<td>558</td>
<td>6.98</td>
<td>7.2</td>
<td>17.8</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FWN_01-8_POL</td>
<td>N701</td>
<td>N502</td>
<td>3.17</td>
<td>85.59</td>
<td>700</td>
<td>6.86</td>
<td>7.29</td>
<td>17.71</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FWN_161_POL</td>
<td>N703</td>
<td>N501</td>
<td>2.15</td>
<td>58.05</td>
<td>793</td>
<td>4.11</td>
<td>12.17</td>
<td>12.83</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FWS_03-250_POL</td>
<td>N702</td>
<td>N501</td>
<td>4.82</td>
<td>130.14</td>
<td>746</td>
<td>9.79</td>
<td>5.11</td>
<td>19.89</td>
</tr>
<tr>
<td>Pleistodontes</td>
<td>FWS_137_POL</td>
<td>N702</td>
<td>N502</td>
<td>4.93</td>
<td>133.11</td>
<td>809</td>
<td>9.23</td>
<td>5.42</td>
<td>19.58</td>
</tr>
</tbody>
</table>

**Appendix B:** Details of the indices used to identify each individual in the pools and the quantities used to dilute the library preparations to equal concentrations so they can be mixed in the same pool.
Population genomics of a fig wasp community

Appendix C – Blobplots

Pleistodontes (ex F. obliqua) – FON_398_POL
*Pleistodontes (ex F. obliqua) – FON_403_POL*
Pleistodontes (ex F. obliqua) – FOS_183_POL1
Population genomics of a fig wasp community

Pleistodentes (ex F. obliqua) – FOS_187_POL
Sycoscapter (ex F. obliqua) – FON_339_SYC
Sycosceptr (ex F. obliqua) – FON_TrinBeach_SYC
Sycoscerter (ex F. obliqua) – FOS_186_SYC
Population genomics of a fig wasp community

*Sycoscapter (ex F. obliqua) – FOS_199_SYC*
Philotrypesis (ex F. obliqua) – FON_340_PHIL
Philotrypesis (ex F. obliqua) – FON_PDoug_PHIL
Philotrypesis (ex F. obliqua) – FOS_186_PHIL
Population genomics of a fig wasp community

*Philotrypesis (ex F. obliqua) – FOS_188_PHIL1*
Sycoscapter long (ex F. rubiginosa) – FRN_274_SYC1
Sycoscapter long (ex F. rubiginosa) – FRN_350_SYC
Sycosapter long (ex F. rubiginosa) – FRS_32_SYC
Sycoscapter long (ex F. rubiginosa) – FRS_397_SYC
Sycosceptr long (ex F. rubiginosa) – FRSYD_89_SYC
Sycosapter long (ex F. rubiginosa) – FRSYD_2873_SYC
Sycoscapter short (ex F. rubiginosa) – FRN_332_SYC
Sycoscapter short (ex F. rubiginosa) – FRS_238_SYC
Population genomics of a fig wasp community

Sycoscapter short (ex F. rubiginosa) – FRS_508_SYC
Population genomics of a fig wasp community

Sycoscapter short (ex F. rubiginosa) – FRSYD_1941_SYC
*Philotrypesis (ex F. rubiginosa) – FRN_310_PHIL*
**Philotrypesis (ex F. rubiginosa) – FRN_407_PHIL**
Philotrypesis (ex F. rubiginosa) – FRS_40_PHIL
**Populatation genomics of a fig wasp community**

*Philotrypesis (ex F. rubiginosa) – FRS_110_PHIL*
Philotrypesis (ex F. rubiginosa) – FRS_247_PHIL
**Philotrypesis (ex F. rubiginosa) – FRS_511_PHIL**
*Philotrypesis (ex F. rubiginosa) – FRSYD_165_PHIL*
Population genomics of a fig wasp community

Philotrypesis (ex F. rubiginosa) – FRSYD_1822_PHIL
Watshamiella (ex F. rubiginosa) – FRN_299_WAT
Population genomics of a fig wasp community

Watshamiella (ex F. rubiginosa) – FRN_337_WAT
Watshamiella (ex F. rubiginosa) – FRS_239_WAT
Population genomics of a fig wasp community

Watshamiella (ex F. rubiginosa) – FRS_514_WAT
Watshamiella (ex F. rubiginosa) – FRSYD_567_WAT
Watshamiella (ex F. rubiginosa) – FRSYD_1905_WAT1
Herodotia (ex F. rubiginosa) – FRN_349 HERO
Herodotia (ex F. rubiginosa) – FRN_372 HERO
Herodotia (ex F. rubiginosa) – FRS_244 HERO
Population genomics of a fig wasp community

Herodotia (ex F. rubiginosa) – FRS_517 HERO
Herodotia (ex F. rubiginosa) – FRSYD_293_HERO
Herodotia (ex F. rubiginosa) – FRSYD_381_HERO
Herodotia (ex F. rubiginosa) – FRSYD_1964_HERO
Population genomics of a fig wasp community

Eukobelea (ex F. rubiginosa) – FRN_310_EUKO
Eukobelea (ex F. rubiginosa) – FRN_406_EUKO
Eukobelea (ex F. rubiginosa) – FRS_41_EUKO
Eukobelea (ex F. rubiginosa) – FRS_239_EUKO

Appendix C: Blobplots for each of the sequenced individuals not included in the main text. The blobs represent contigs, the size of each corresponding to contig length, and are coloured according to phylum designation. The numbers in each key correspond to: the total number of contigs designated to that phylum, the total length of the contigs (span) designated to that phylum, the average length of the contigs designated to that phylum. The phyla are ordered by the total contig length, from highest to lowest.
Population genomics of a fig wasp community
Appendix D – PSMC parameter choice

P. nigriventris PSMC plot (-p "45°2 -t10")

P. nigriventris PSMC plot (-p "45°2 -t20")
Population genomics of a fig wasp community
Population genomics of a fig wasp community

**P. nigriventris PSMC plot (\(-p \, ^145*2+4+6\, -t10\))**

![Graph 1](image1)

**P. nigriventris PSMC plot (\(-p \, ^145*2+4+6\, -t20\))**

![Graph 2](image2)
Appendix D: PSMC analyses testing a range of parameters (varying the –p and –t parameters).