Chapter 1

Introduction

> "By morning the wind had brought the locusts; they invaded all Egypt and settled down in every area of the country in great numbers. Never before had there been such a plague of locusts, nor will there ever be again. They covered all the ground until it was black. They devoured all that was left after the hail—everything growing in the fields and the fruit on the trees. Nothing green remained on tree or plant in all the land of Egypt."
>
> Exodus 10: 13-14

Population genetics and phylogeography have been shown to be indispensable in the understanding of species demographics (Avise *et al.* 1987, Knowles & Maddison 2002, Knowles 2004). The purpose of this dissertation is to apply population genetics principles to gain a better understanding of demographic processes in the African Wild Silk Moth (*Gonometa postica*). Therefore, the purpose of this introductory chapter is firstly, to review the known biology of the African Wild Silk Moth and secondly, to introduce the necessary population genetic principles and methods that will be used for demographic inference in later chapters.

The African Wild Silk Moth is a species that is currently of economic interest in southern Africa. Both this species and its sister species, *G. rufobrunnea*, have been shown to possess a silk fibre of exceptional quality (Freddi *et al.* 1993, Akai *et al.* 1997). In this respect the initiation of an African Wild Silk Industry in southern Africa has been proposed as a potential means of poverty alleviation in rural southern Africa. However, a consistent complaint from small-scale cottage industries that currently utilize *Gonometa* silk is the insufficient supply of cocoons. Since the industry currently only utilizes cocoons from which the adult moths have already emerged, there is little or no effect of harvesting on the population dynamics of the species. Rather, the insufficient supply of cocoons is directly related to the complex population cycles experienced by the species. The species is characterized by two generations per year, the first starting in September-October when adult moths emerge from cocoons. Adult moths emerge without feeding mouthparts and survive for three to five days (maximum nine days, Hartland-Rowe 1992) during which breeding occurs. Eggs are laid, larvae emerge and pass through six instar larval stages in approximately five weeks, after which the larvae construct cocoons, pupate and enter a period of diapause. This period of diapause either carries through to the following September when adults emerge, or is broken in February with adult emergence and an additional population cycle. Typically, this second generation comprises between 12-50% of the first generation (Hartland-Rowe 1992), and culminates in pupae that emerge as adults in September. *G. postica* experiences large inter-annual population size fluctuations (Veldtman 2004), though it is uncertain which factors contribute to the cyclical nature of this species.

In order to understand eruptions of this species several questions need to be addressed. Firstly, the influence of climatic factors on the incidence of population eruptions should be evaluated. It has been hypothesized that *G. postica* eruptions follow periods of drought, where the rates of larval parasitism are reduced during these times, thus allowing the normally heavily-parasitised larvae (Veldtman 2004) to reach eruptive proportions (Hartland-Rowe 1992). Secondly, the interaction between host plant, *Acacia erioloba,* phenology and *G. postica* is unknown. *A. erioloba* experiences a leaf-flush in August prior to the emergence of adult moths (Smit 1999). The timing of this leaf-flush in relation to rainfall, and the timing of *G. postica* emergence, is crucial for the understanding of eruptions in this species. Related to the phenology of the host plant, it is necessary to determine the quantity and quality of foliage required for larvae to complete development, pupation and emergence as adult moths. This interaction between climate, host-plant and *G. postica* would be crucial for the understanding of complex population dynamics in this species. Ideally, a long-term population dynamics programme should be initiated that evaluates the effect of both exogenous and endogenous factors in determining local annual population sizes of this species. The scale, however, at which such a study should be conducted is uncertain since there are currently no dispersal estimates available, and therefore it is uncertain what constitutes a population in this species. Such knowledge of dispersal ability will further enhance the interpretation of the temporal occurrence of eruptions, where eruptions in later years are sourced from nearby eruptions in preceding years. The purpose of this dissertation is to estimate the degree of genetic connectivity between eruptions within and between years, using spatial population genetic analysis. Such an understanding, of dispersal ability in this species, will allow the planning of harvesting strategies and potentially allow the incorporation of a dispersal parameter into predictive distribution modeling that is planned for future research.

Several species exhibit complex population cycles and large fluctuations in both density and population size (Finerty 1979, Bjornstad *et al.* 2002, Tallmon *et al.* 2002, Turchin 2003). Turchin (2003) has reviewed the dynamics of such complex population cycles and notes that such cycles are the result of the interaction between endogenous and exogenous factors. These factors encompass environmental effects, including climatic factors, population-specific effects such as density dependence and inter-species interactions, such as predator-prey relations. Although the ecological literature of complex population cycles is extensive, comparatively few genetic studies have been conducted. Population genetic analysis of snowshoe hare (*Lepus americanus*) have attributed the observed spatial patterns of genetic variation to a stepping stone model of gene flow influenced by density cycles, where local bottleneck populations expand to previously unsuitable habitat and thus homogenize genetic diversity across the distribution (Burton *et al.* 2002). The collared lemming, another cyclic species, also has a spatial genetic pattern characterized by very little population structuring. In this species, the inferred high levels of gene flow are attributed to long-distance dispersal events (Ehrich *et al.* 2002). Similarly, deviation from an isolation by distance model in spatial genetic structure in the butterfly *Aglais urticae* is attributed to high movement rates, occasional long-distance migration and rare extinction/recolonisation events (Vandewoestine *et al.* 1999). Northeastern Australian rabbit populations also show large degrees of population size fluctuations, yet differ in degree of stochasticity between the arid west and semiarid east (Fuller *et al.* 1997). Spatial genetic patterns in this species corroborate the observation of a high degree of gene flow inferred from population genetic data of cyclical species in that the western populations exhibit reduced levels of structure versus that of the eastern populations, which have fewer stochastic fluctuations in population size (Fuller *et al.* 1997). This general result of very little population structuring in cyclical species may be worthy of further investigation (Burton *et al.* 2002). Theoretically, one might expect a higher degree of population sub-structuring in species that exhibit population cycles (Wright 1940), due to increased probability for different demes to become fixed for alternate alleles, under random genetic drift during periods of small population size. However, this effect is most likely dependent on the levels of dispersal and whether fixation of alleles at particular demes can be removed due to movement of alleles between demes in years of population size expansion. The effects of extinction and recolonisation on spatial genetic pattern have been evaluated for species with metapopulation structure (Wade & McCauley 1988, Whitlock & McCauley 1990, Ibrahim *et al.* 2000, Ibrahim 2001). These results in general indicated that the effect of population turnover on genetic differentiation is dependent on the number of individuals colonizing a deme relative to the number of recurrent migrants between demes (Whitlock & McCauley 1990, Ibrahim 2001). This is intuitive for metapopulations since low numbers of founders are likely to produce greater genetic structure, as is colonization from single versus multiple demes.

Recurrent migrations will furthermore tend to homogenize genetic diversity given high levels of migration. These results generally appear to hold for metapopulations, yet some species do not have obvious metapopulation structure and simply exist as continuous populations where neighbourhood sizes fluctuate as a result of local changes in density. Dispersal in this instance does not occur between spatially defined demes, but are rather effected as an individual dispersal distance in a continuously distributed isolation by distance model. The effects of population size variations in such species are likely to be different, and thus are explored within the context of this dissertation.

Inference of population demography from genetic data

In order to address the dispersal ability of *G. postica* through spatial genetic analysis it is necessary to use the currently available analysis approaches for spatial genetic data. The purpose of the following section is to introduce the available methods. This review of spatial genetic analysis methods is by no means exhaustive. Rather it is a personal reflection on the development of the field and is biased towards simulation and coalescent modeling approaches. Furthermore, although I have utilized spatial autocorrelation analyses methods in the subsequent chapters I have not covered these here. I feel that spatial autocorrelation approaches do not contribute to the development of custom demographic analysis models that I personally believe is the future of spatial population genetic analysis.

The inference of demographic parameters from population genetic data is a field that has grown rapidly in past years. This field, originally termed phylogeography (Avise *et al.* 1987) originated in the 1980's and has enjoyed a long tradition of gathering spatial genetic data, and subsequently inferring processes from the correlation of such data with landscape, or historical geographical/climatic features and events (see reviews by Avise *et al.* 1987, Avise 2000). This spatial pattern matching, however, is fraught with ideological and theoretical problems, the most notorious being the inference of complex demographic processes from a single gene tree. As such phylogeography and analyses of spatial genetic data has moved from a pattern-based descriptive science to one that involves the statistical testing of alternate hypotheses against the observed genetic data. The rapid increase in computer power in the last decade has fueled this development of statistical phylogeography, which has been the subject of a recent special issue of *Molecular Ecology* (Volume 13, 2004). In this introduction I briefly review the history of spatial population genetic analysis and introduce the current advances in statistical phylogeography, and demographic inference. The purpose of this discussion is to provide a framework for the analysis of spatial population genetic data collected from the focal species of this dissertation.

Summary statistic approaches

Probably the most important consideration for any analysis technique is the particular demographic/migration model on which a particular technique is based. Three models are common in the population genetics literature, the island model (Wright 1940, Crowe 1986), the stepping-stone model (Kimura & Weiss 1964, Nagylaki 1982), and the isolation by distance model (Wright 1943), where a suite of summary statistics characterizes each. Historically, population genetics, and thus inference of demographic parameters such as migration, has been strongly based on summary statistics. Wright (1951) devised an approach to partitioning genetic variation in a subdivided population based on the island model of migration, termed Fixation indices or *F*-statistics. The calculation of F_{ST} has dominated the population genetics literature, and is simply the variance in allele frequencies across populations (V_a) standardized by the mean allele frequency, ρ.

$$F_{ST} = \frac{V_a}{\rho(1-\rho)} \qquad \text{Wright (1951)}$$

Several methods are available to estimate F_{ST} (Wright 1951, Weir & Cockerham 1984, Nei 1987), yet most make use of the relationship above. However, Slatkin (1985, 1987) has suggested an approach based on the distribution and frequency of rare alleles, and Barton and Slatkin (1986) have found these alternate measures to be consistent over a wide range of assumptions of population structure, selection and mutation. Recent reviews of F_{ST}, as a measure of population differentiation, should be consulted for further discussion (Weir & Hill 2002, Excoffier 2003).

Of particular importance in calculating F_{ST} are the underlying assumptions of the model, i.e. an island migration model in a Wright-Fisher population of constant size.

The principal aim of calculating F_{ST} is the inference of both effective population size (N_e), and migration rates between demes (m);

$$N_e = \frac{Nd}{1 - F_{ST}} \qquad \text{Wright (1931)}$$

$$F_{ST} = \frac{1}{4Nm + 1} \qquad \text{Wright (1931)}$$

where N = population size, d = the number of demes and m = migration rate. However, considerations of the assumptions, in terms of underlying population structure, of such calculations are paramount. Indeed, Whitlock (2004) has considered the application of the above estimators of N_e and *4Nm* applied to metapopulations. A critical assumption in the above model is that of no variance in reproductive success among demes. Since F_{ST} typically takes on values between 0 and 1, the above estimator of N_e gives the nonsensical result of N_e always being greater than *Nd*, the product of population size per deme, *N*, and number of demes, *d* (Whitlock 2004). This result is contrary to what is expected is natural metapopulations, where large variances in reproductive success among demes is expected, and thus highlights the importance of considering the assumptions behind a particular model when analyzing and interpreting data. Several theoretical population models have been developed for the estimation of *F*-statistics and the analysis of population structure, including extinction-recolonisation metapopulation models (Whitlock & McCauley 1990), source-sink models (Gaggiotti 1996), and stepping-stone models (Kimura & Weiss 1964). However, since the purpose of many population genetic studies is to infer the underlying genetic structure of the focal species, the suitable model is not known *a priori*. Thus the development of statistical procedures that can simultaneously estimate the underlying genetic structure of a population, and demographic parameters, is a central challenge in population genetics (Excoffier 2003).

Given a continuous population the isolation by distance model (Wright 1943) is most appropriate, where summary statistics of interest are those concerned with neighbourhood size. Neighbourhood size essentially represents the number of individuals an individual would encounter within its lifetime, and is dependent on density (D) and the standard deviation of the distribution of dispersal distances (σ).

The size of a neighbourhood (N_b) amounts to the number of individuals in a circle with a radius twice the standard deviation of dispersal distances (2σ). Thus,

$$N_b = 4\pi D\sigma^2 \qquad \text{(Wright 1943)}$$

Rousset (1997, 2000) has developed methods for the estimation of neighbourhood size from pairwise calculations of F_{ST} between demes and between individuals. Typically, the method involves the calculation of pairwise genetic distances, and subsequent plotting against the natural logarithm of distance. The inverse of the slope of the regression provides an estimate of neighbourhood size (Rousset 1997, 2000).

Another summary statistic of importance is the calculation of θ, the composite estimate of population size (N) and mutation rate (μ); $2N\mu$ in haploids and $4N\mu$ in diploids. The need for estimating θ arises from the fact that neither of the two measures that comprise the parameter can be estimated independently from population genetic or sequence data without prior information on mutation rates or effective population size. Mutation rate, in particular is notoriously difficult to estimate from genetic data due to the occurrence of back mutations. Back mutations occur in DNA sequence data since mutation rate variation typically follows a distribution with few sites of high mutation rate and many sites of low mutation rates (Yang 1996). Several estimators of θ are evident in the literature: the expected number of alleles in an infinite-allele model (Ewens 1972), the number of segregating sites in an infinite-site model (Watterson 1975) or nucleotide diversity, the mean number of pairwise differences (Tajima 1983). The parameter θ also proves useful in maximum-likelihood inference of demographic parameters given the observed genetic data.

Furthermore, the relationship between different estimators of θ forms the basis for Tajima's *D* (Tajima 1989), the test of neutrality of mutations. Tajima's *D* is based on the premise that in a gene under selection, θ estimated from segregating sites will be substantially greater than θ estimated from nucleotide diversity, since rare mutations that are selected against are down-weighted in the calculation of the latter. Fu and Li (1993) have further developed neutrality tests based on the observation that purifying selection is evident as an excess of mutations at the tips of a gene genealogy. Since these tests compare mutations in the recent past to mutations in the distant past, the use of an outgroup from a closely related species is recommended (Fu & Li 1993). Although neutrality tests have been criticized for lack of statistical power (Simonsen *et al.* 1995), a central problem relating to neutrality tests is rather the difficulty in distinguishing selection for a particular allele versus the demographic event of rapid population growth. Both of these processes generate the same genetic signal (Tajima 1989), evident as a star-like pattern in a gene tree, or unimodal distribution of pairwise genetic differences, or mismatch distributions (Rogers & Harpending 1992). The use of multiple loci is thus crucial in any population genetics study, where selection will be evident at only those loci under selection, whereas demographic histories will be evident in the genetic patterns of all loci.

An extension to the analysis of mismatch distributions is the development of methods to analyse a spatial range expansion versus that of simply an increase in population size (Ray *et al.* 2003, Excoffier 2004). Initially, a coalescent simulation combined with a demographic model of spatial expansion was used to observe the effects of spatial expansions on intra-deme molecular diversity in a structured population model (Ray *et al.* 2003). The results in general indicated that under low levels of migration between demes a spatial demographic expansion generated multimodal mismatch distributions. In contrast, large levels of migration between demes generated a pattern that was indistinguishable from a structured population that had always been exchanging a large number of migrants (Ray *et al.* 2003). Excoffier (2004) further utilized these simulation results to derive an analytical expression of F_{ST} given a structured population that has undergone a recent spatial demographic expansion. The process of simulating a demographic process and observing the effects on patterns of genetic diversity and subsequently devising an analytical algorithm is conducive to the advancement of population genetic theory. In addition, the results from a simulation model, and the data generated, can be subsequently input into the analytical algorithm such that its power to detect demographic processes can be evaluated.

Model-based approaches

Model-based approaches to analyzing and interpreting spatial genetic patterns are subdivided into three general approaches, comparative simulation modeling, analytical-based inference and simulation-model-based inference. As is evident in the example above (Ray *et al.* 2003) comparative simulation modeling provides a means to understand the effects that a particular demographic process may have on inference using summary statistics. Typically, comparative simulation modeling would comprise the repeated simulation of a demographic event, and the repeated estimation of a summary statistic of interest. General conclusions can thus be drawn regarding the effect of the demographic event on the calculation of the statistic of interest. Such simulations can either be effected forward-through-time where every individual in the population is simulated, and a sub-sample is drawn for the calculation of summary statistics; or backwards-through-time where only the genealogical history of the sampled alleles need to be simulated. Such backwards-through-time models, based on coalescent theory, have become popular due to their simplicity and mathematical tractability. The neutral coalescent, based on a panmictic Wright-Fisher population of constant population size, with no selection and no recombination, simply states that of all possible events that could happen to a sample of *n* alleles one generation back in time, only two are important: either all *n* alleles have distinct parents, or two alleles in the sample share a common ancestor (Wakeley 2004). This process continues from the samples observed in the present, backwards through time, until all samples (lineages) have coalesced to a single common ancestor. The result is a genealogy of the samples characterized by *n-1* coalescent events, and a distribution of times to each coalescent event (branch lengths). Kingman (1982) showed that the probability that *n* alleles are reduced to *n-1* alleles in the previous generation, in a total population size *N*, is given by

$$P_n = \frac{n(n-1)}{4N} \qquad \text{Kingman (1982)}$$

and thus the estimated time for *n* alleles to be reduced to *n-1* alleles is given by

$$E(T_n) = \frac{4N}{n(n-1)} \qquad \text{Kingman (1982)}$$

Therefore, since the genealogy comprises *n-1* coalescent events the total time to coalescence is given by

$$E(t) = \sum_{i=2}^{n} E(T_i) = 4N(1 - \frac{1}{n})$$

Per generation coalescent demographic simulation is a two-step process. Firstly, a random genealogy is simulated under the demographic model of interest, followed by the scattering of mutations onto the genealogy given the branch lengths and a mutation rate. The demographic model of population size fluctuations for example, could be implemented through varying *N* in the equations above at each generation to calculate the probability of coalescence. These probabilities are used to construct random genealogies, by drawing a random number at each generation. Coalescent events occur when a random number is greater than the probability of coalescence at the particular generation and is subsequently recorded as a coalescent event in the genealogy. Mutations are scattered forward through time on each simulated genealogy according to the branch lengths and mutation rates, and thus genetic data is generated at the tips of the genealogy. The process is repeated thousands of times, each time generating data and calculating a summary statistic of interest. Distributions of the summary statistics of interest can then be observed to infer the effect of the demographic process simulated. The result is an understanding of the potential variance in genetic data under a particular demographic model, given that the gene sorting process within populations is stochastic. Leblois *et al.* (2004) have used such an approach to determine the effects of temporal changes in density and dispersal on the inference of neighbourhood size in continuous populations.

In some cases analytical results are available for statistics of interest. For example, the analytical formulation of the number of segregating sites in a sample *n* can be used to make a maximum likelihood estimate of θ. However, this represents a point estimate of θ, whereas a degree of error is often required. Obtaining a variance for such estimation is not yet possible (Wakeley 2004). Furthermore, the variance of an estimator is only useful when the errors are normally distributed, or when the distribution is known and symmetric, a characteristic atypical of genetic data

(Wakeley 2004). Thus the analytical-based inference of demographic parameters given genetic data is limited and a simulation-based model of inference is required.

Simulation-based inference is dependent on maximum-likelihood (Edwards 1972) and Bayesian methods (Bayes 1763), and thus these are first introduced. Likelihood-based inference has the aim of evaluating the likelihood of a particular parameter given the observed data. Thus the calculation of the probability $P(D \mid \psi)$, of observing data, D, given the parameters, ψ, of a particular model is performed. The parameters of the model typically include the genealogy of the sampled alleles, the population size-mutation rate composite, θ, migration rates, and rates of population size increase or decrease. Bayesian inference procedures also utilize likelihood in the calculation of a posterior distribution for parameters, ψ, but further allow the incorporation of prior knowledge of the system under consideration. Such prior knowledge may take the form of a direct-estimate of migration rates between demes, or some prior belief of the effective population size. The posterior distribution is given by

$$P(\psi \mid D) = \frac{P(D \mid \psi)P(\psi)}{P(D)} \qquad \text{Bayes (1763), Edwards (1972)}$$

where $P(\psi)$ is the distribution of a prior. Priors in population genetic applications are typically uniform and given sufficient data the posterior distribution is dominated by the likelihood, such that the choice of prior has little effect on the conclusions drawn (Stephens 2003). In the following discussion of inference, however, I will concentrate on likelihood-based inference since the procedure of estimation in Bayesian inference is similar.

Coalescent-based likelihood inference is based on the distribution of genealogies, τ, given the observed data. Associated with each genealogy are parameters, such as branches, or mutations on the branches. Given that τ is known, one could calculate the probability of the data given the parameters of the model, i.e. $P(D \mid \tau, \psi)$. However, τ is unknown and as such one needs to calculate the likelihood of a particular parameter, or suite of parameters, by summing over all possible genealogies. This is performed since a single data set can be obtained from many different genealogical histories. As such, the likelihood of the parameter is the sum of all probabilities for all potential genealogies, given the parameters, ψ

$$L(\psi) = P(D \mid \psi) = \sum_{\tau} P(D \mid \tau, \psi) P(\tau \mid \psi) \qquad \text{Felsenstein (1973, 1988)}$$

The number of possible genealogies increases rapidly with the number of samples, for three samples there are only three possible genealogies, for 10 samples 2571912000 genealogies and for 100 samples, 1.37 x 10^{284} genealogies (Felsenstein 2004). Thus, summing over all possible genealogies becomes computationally unfeasible given the sample sizes of typical population genetics studies. However, a glimmer of hope resides in the observation that the proportion of genealogies that contribute significantly to the sum across all genealogies is less than one in a million (Stephens 2003). This forms the basis of inference techniques, such as Monte Carlo integration (Hastings 1970), which have the purpose of estimating the likelihood surface through estimating the aforementioned likelihoods over parameter space. In summary, MCMC focuses the calculation of likelihoods in genealogy space where the genealogies that contribute most to the likelihood reside. Thus the amount of computation is reduced and likelihood surfaces can be approximated. Typically, the process involves starting at a particular point in genealogy space, evaluating the probabilities in the likelihood function above, proposing an alteration to the genealogy such that a move in genealogy space is suggested, and deciding whether to move to the suggested point based on the ratio of the current probability to that of the proposed point. These movements through genealogy space are typically referred to as chains. Coalescence and importance sampling are extensively reviewed in both Stephens (2003) and Felsenstein (2004) and thus will not be considered further.

Currently, the coalescent-based inference of demographic parameters is restricted to only a few migration models and demographic scenarios. These scenarios include the estimation of (i) gene flow in structured populations (Beerli & Felsenstein 1999, 2001), (ii) population growth/decline in both unstructured (Kuhner *et al.* 1998) and structured populations (Kuhner *et al.* 2004), (iii) divergence time of two populations that exchange/exchanged migrants (Nielsen & Wakeley 2001) and (iv) recombination rates in structured/unstructured populations (Fearnhead and Donnelly 2001, Kuhner *et al.* 2004). There is a potential to include more demographic scenarios in coalescent-based models, such as fluctuating population size in structured populations, and indeed Knowles and Maddison (2002) and Knowles (2004) have noted that custom-development of species-specific models for demographic inference will be the future of phylogeographic analysis. Thus far I have only considered maximum likelihood and Bayesian inference models that make full use of the sequence or allelic data provided. However, there is a trend in the population genetics literature towards maximum-likelihood or Bayesian inference conditional on summary statistics (Tavaré *et al.* 1997, Weiss & von Haeseler 1998, Beaumont *et al.* 2002). These methods differ from the inference process described above in that at each step in the chain, the probability of the data given the parameters $P(D \mid \psi)$ is replaced with the probability of some summary statistic, *k*, given the parameters, i.e. $P(k \mid \psi)$. Given that the summary statistic is a sufficient representation of the data, the evaluation of these likelihoods is computationally faster than full-data methods. These methods may allow the development of multi-parameter custom likelihood models for genetic data analysis.

Dissertation outline

Chapter 1: Introduction and literature review

Chapter 2: Temporal and spatial distribution of African Wild Silk Moth, *Gonometa postica*, eruptions in southern Africa

This chapter presents three years of presence/absence distribution data. The potential cause of large-scale temporal changes in the distribution of eruptions is discussed and recommendations are made with regard to future temporal data collection, and climatic modeling.

Chapter 3: Characterisation of six microsatellite loci in the African Wild Silk Moth (*Gonometa postica*, Lasiocampidae)

Species-specific microsatellite markers were developed for *Gonometa postica* using an AFLP-based enrichment protocol (Zane *et al.* 2002). The six loci are reported along with estimates of Hardy-Weinberg and linkage-disequilibrium. This chapter is in press with *Molecular Ecology Notes*.

Chapter 4: The effect of large annual population size fluctuations on spatial genetic pattern in the continuously distributed African Wild Silk Moth (*Gonometa postica*)

The results of a single years analysis of microsatellite and mtDNA genetic data are presented in this chapter. The results are peculiar in that a species that has low dispersal ability and only lives for a few days appears to have little evidence for isolation by distance in the data. Simulations are subsequently used in an effort to understand the effects of population size fluctuations on spatial genetic pattern in a continuously distributed species. The analysis in this chapter falls within the section 'comparative simulation modeling' as described above.

Chapter 5: Temporal and spatial genetic patterns in the African Wild Silk Moth (*Gonometa postica*) and implications for cyclical population dynamics

The population genetic results from three years of successive sampling are presented in this chapter. The analysis is focused on estimating the degree of population size changes and detecting whether there may be weak signals of migration in the data. Standard maximum-likelihood techniques are used, and the implications of the assumptions of these techniques in analyzing the data are discussed.

Chapter 6: CoalFace: a graphical user interface program for the simulation of coalescence

Coalescent analysis is at the forefront of current population genetics, and thus in the preparation for a local workshop on phylogeographic analysis I developed a tool to teach the applications of coalescent theory. This chapter describes the software and its applicability in teaching population genetics and phylogeography.

Chapter 7: Conclusions

Appendix I: A population genetics pedigree perspective on the transmission of *Helicobacter pylori*

During my PhD I collaborated on a project to determine the mode of transmission of *Helicobacter pylori* using gene sequences and an extensive pedigree. The study presented a unique opportunity to investigate transmission of this bacterium, due to the abnormally high prevalence in a local South African community and the extensive sampling. I developed a simulation model, in collaboration with co-authors, which contrasted different modes of transmission and their effect on observed genetic data. This chapter is included as an Appendix here since although the species, and question are different from my principal dissertation research, the method of using simulation modeling for inference is common.

Appendix II: LatticeFlucII source code

The C source code for the simulation model used in chapter 4 is presented.

Appendix III: CoalFace source code

The kylix/Delphi source code for the CoalFace program presented in chapter 6 is presented.

General Notes

Please note that all chapters (except 1 and 7) are written as manuscripts that have been or will be submitted for publication. Since all chapters have supervisors or collaborators as co-authors, I refer to the work being done by us, and not exclusively by myself. However, the research presented in this dissertation is entirely my own thought and execution, with useful discussions with my supervisors. The work presented in Appendix I, however, was a combined effort between Michael Cunningham, myself and other co-authors, and thus I have not included it in the main part of the dissertation. However, I still have first authorship for developing the simulation model and writing the manuscript. Since each chapter is in manuscript format, each has a list of authors as will appear in published form. In addition, the reference lists occur at the end of each chapter rather than at the end of the dissertation. Due to this format, there may be instances of duplication across chapters.

References

Akai H, Nakatomi R, Kioko E, Raina SK (1997) Fine structure of cocoon and cocoon filament from African *Gonometa* silkmoth (Lasiocampidae). *Int. J. Wild Silkmoth & Silk*, **3**, 15-22.

Avise JC (2000) Phylogeography: The history and formation of species. Harvard University Press, Harvard.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA & Saunders NC. (1987) Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology & Systematics*, **18**, 489-522.

Barton NH & Slatkin M (1986) A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity*, **56**, 409-416.

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, **53**, 370-418. Reprinted In: Studies in the History of probability and statistics. IX. Thomas Bayes' essay towards solving a problem in the doctrine of chances. *Biometrika*, **45**, 293-315.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population Genetics, *Genetics*, **162**, 2025-2035.

Beerli P & Felsenstein J. (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763-773.

Beerli P & Felsenstein J. (2001) Maximum-likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences, USA*, **98**, 4563-4568.

Bjornstad ON, Peltonen M, Liebhold AM & Balstensweiler W (2002) Waves of larch budmoth outbreaks in the European Alps. *Science*, **298**, 1020-1023.

Burton C, Krebs CJ, Taylor EB (2002) Population genetic structure of the cyclic snowshoe hare (*Lepus americanus*) in southwestern Yukon, Canada. *Molecular Ecology*, **11**, 1689-1701.

Crow JF (1986) *Basic Concepts in Population, Quantitative and Evolutionary Genetics*. WH Freeman, San Francisco.

Edwards ACF (1972) Likelihood: expanded edition. The Johns Hopkins University Press, London.

Ehrich D, Jorde PE, Krebs CJ, Kenney AJ, Stacy JE, Stenseth NC (2001) Spatial structure of lemming populations (*Dicrostonyx groenlandicus*) fluctuating in density. *Molecular Ecology*, **10**, 481-495.

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87-112.

Excoffier L (2003) Analysis of population subdivision. In: Handbook of Statistical Genetics, 2nd Edition (eds. DJ Balding, M Bishop, C Cannings), pp 713-750, John Wiley & Sons, Ltd.

Excoffier L (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology*, **13**, 853-864.

Fearnhead P & Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299-1318.

Felsenstein J (1973) Maximum likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, **25**, 471-492.

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics*, **22**, 521-565.

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates Inc. Massachusetts.

Finerty JP (1979) Cycles in Canadian Lynx. *American Naturalist*, **114**, 453-455.

Freddi G, Bianchi Svilokos A, Ishikawa H, Tsukada M (1993) Chemical composition and physical properties of *Gonometa rufobrunnea* silk. *Journal of Applied Polymer Science*, **48**, 99-106.

Fu Y-X & Li W-H (1993) Statistical tests of Neutrality of Mutations. *Genetics*, **133**, 693-709.

Fuller SJ, Wilson JC & Mather PB (1997) Patterns of differentiation among wild rabbit populations *Oryctolagus cuniculus* L. in arid and semiarid ecosystems of north-eastern Australia. *Molecular Ecology*, **6**, 145-153.

Gaggiotti OE (1996) Population genetic models of source-sink metapopulations. *Theoretical Population Biology*, **50**, 178-208.

Hartland-Rowe R (1992) The Biology of the wild silkmoth *Gonometa rufobrunnea* Aurivillius (Lasiocampidae) in northeastern Botswana, with some comments on its potential as a source of wild silk. *Botswana Notes and Records*, **24**, 123-133.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Ibrahim KM (2001) Plague dynamics and population genetics of the desert locust: can turnover during recession maintain population genetic structure. *Molecular Ecology*, **10**. 581-591.

Ibrahim KM, Nichols RA, Hewitt GM (1996) Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity*, **77**, 282-291.

Kimura M & Weiss GH (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, **61**, 763-771.

Kingman JFC (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235-248.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623-2635.

Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**, 1-10.

Kuhner MJ, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on coalescent. *Genetics*, **149**, 429-434.

Kuhner MK, Yamato J, Beerli P, Smith LP, Rynes E, Walkup E, Li C, Sloan J, Colacurcio P, Felsenstein J (2004) LAMARC v 1.2.1. University of Washington, http://evolution.gs.washington.edu/lamarc.html.

Leblois R, Rousset & Estoup A. (2004) Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics*, **166**, 1081-1092.

Nagylaki T (1982) Geographical invariance in population genetics. *Journal of Theoretical Biology*, **99**, 159-172.

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York, NY, USA.

Nielsen R, Wakeley JW (2001) Distinguishing Migration from Isolation: an MCMC Approach. *Genetics*, **158**, 885-896.

Ray N, Currat M, Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution*, **20**, 76-86.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552-569.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*,**145**, 1219-1228.

Rousset F (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology*, **13**, 58-62.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413-429.

Slatkin M (1985) Rare alleles as indicators of gene flow. *Evolution*, **39**, 53-65.

Slatkin M (1987) Gene flow and the geographic structure of natural populations, *Science*, **236**, 787-792.

Smit N (1999) A Guide to the Acacias of South Africa. Briza, Pretoria.

Stephens M. (2003) Inference under the coalescent. In: *Handbook of Statistical Genetics, 2nd Edition.* (eds. DJ Balding, M Bishop, C Cannings), pp 636-661, John Wiley & Sons, Ltd.

Tajima F (1983) Evolutionary relationships of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585-595.

Tallmon DA, Draheim HM, Mills L & Allendorf FW (2002) Insights into recently fragmented vole populations from combined genetic and demographic data. *Molecular Ecology*, **11**, 699-709.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505-518.

Turchin (2003) Complex Population Dynamics: a theoretical/empirical synthesis. *Monographs in Population Biology*, **35**. Princeton University Press, Princeton.

Vandewoestine S, Neve G, Baguette (1999) Spatial and temporal population genetic structure of the butterfly Aglais urticae L. (Lepidoptera, Nymphalidae). *Molecular Ecology*, **8**, 1539-1543.

Veldtman R (2004) The ecology of southern African wild silk moths (*Gonometa* species, Lepidoptera: Lasiocampidae): consequences for their sustainable use. University of Pretoria PhD thesis.

Wade MJ & McCauley DE (1988) Extinction and recolonisation: their effects on the genetic differentiation of local populations. *Evolution*, **42**, 995-1005.

Wakeley J (2004) Metapopulations and coalescent theory. In: *Ecology, Genetics and Evolution of Metapopulations* (eds. I. Hanski & O. E. Gaggiotti), pp 175-198. Elsevier Academic Press, London.

Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256-276.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.

Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics*, **36**, 721-750.

Weiss G, von Haeseler (1998) Inference of population history using a likelihood approach. *Genetics*, 149, 1539-1546.

Whitlock MC & McCauley (1990) Some population genetic consequences of colony formation and extinction: genetic correlations with founding groups. *Evolution*, **44**, 1717-1724.

Whitlock MC (2004) Selection and drift in metapopulations. In: *Ecology, Genetics and Evolution of Metapopulations* (eds. I. Hanski & O. E. Gaggiotti), pp 153-174. Elsevier Academic Press, London.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, 16, 97-159.

Wright S (1940) Breeding structure of populations in relation to speciation. *American Naturalist*, **74**, 232-248.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114-138.

Wright S (1951) The genetical structure of populations. *Ann. Eugen*., **15**, 323-354.

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution*, **11**, 367-372.